Comparative and functional analysis of alternative splicing in eukaryotic genomes

Lu Chen

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

January 2012

Table of Contents

Acknowledgements

First I would like to thank my first supervisor Dr. Araxi Urrutia, for the opportunity to work in her laboratory, and her guidance and enormous support, and my second supervisor Prof. Laurence Hurst for his guidance, support and helpful discussion. I really enjoyed my time at Bath for I have been acquiring the skills, academic attitude and way of thinking from you both.

Next I wish to thank the members of the Araxi's and Laurence's Labs, both past and present. Special thanks to Jaime Tovar for all his help, discussion and contribution to our work. I also would like to thank Stephen Bush, Claudia Weber, Marina Angelopoulou, Nina Ockendon, Wei Wang, Guangzhong Wang, Toby Warnecke and Cathy Pink for their help and discussion throughout my PhD. I also want to express my thanks to my friends Yu Qiu, YanJun Huang, Niancai Cheng and many others. I definitely had a good time with you guys. I also would like to add my special thanks to Humberto Gutierrez for reviewing my manuscripts and giving suggestion.

I could not have come this far without the constant support and love from my father, mother and sister. I cannot thank you enough for your encouragement, understanding and faith in me. I also like to thank my nephew as a source of inspiration and joy during my PhD. Last but not least, I am very thankful for the support, encouragement and discussion from Jingwen Lin.

Abbreviations

3’ splice site (3'ss)

5’ splice site (5'ss)

Alternative splicing (AS)

Adenovirus 2 (Ad2)

Complementary DNA (cDNA)

Eukaryote Comprehensive & Comparable Alternative Splicing Events Database (ECCASED)

Exonic splicing enhancer (ESE)

Exonic splicing silencer (ESS)

Expressed sequence tag (EST)

Gene duplication (GD)

Gene expression (GE)

Gene family size (GFS)

Genomic Mapping and Alignment Program (GMAP)

Heterogeneous nuclear ribonucleoproteins (hnRNP)

Intronic splicing enhancer (ISE)

Intronic splicing silencer (ISS)

Message RNA (mRNA)

Next generation sequencing (NGS)

Nonsense-mediate decay (NMD)

Open Reading Frame (ORF)

Polyadenylation (poly-A)

Precursor mRNA (Pre-mRNA)

Reverse transcription polymerase chain reaction (RT-PCR)

Serial analysis of gene expression (SAGE)

Serine/arginine-rich proteins (SR proteins)

Small nuclear ribonucleoprotein (snRNP)

Transposable element (TE)

Abstract

Alternative splicing (AS) is a common post-transcriptional process in eukaryotic organisms, by which multiple distinct functional transcripts are produced from a single gene. Because of its potential role in expanding transcript diversity, interest in alternative splicing has been increasing over the last decade, ever since the release of the human genome draft showed it contained little more than the number of genes of a worm. Although recent studies have shown that 94% human multi-exon genes undergo AS while aberrant AS may cause disease or cancer, evolution of AS in eukaryotic genomes remains largely unexplored mainly due to the lack of comparable AS estimates. In this thesis I built a Eukaryote Comprehensive & Comparable Alternative Splicing Events Database (ECCASED) based on the analyses of over 30 million Expressed Sequence Tag (ESTs) for 114 eukaryotic genomes, including protists (22), plants (20), fungi (23), metazoan (non-vertebrates, 29) and vertebrates (20). Using this database, I addressed two main questions: 1) How does alternative splicing relate to gene duplication (GD) as an alternative mechanism to increase transcript diversity? and 2) What is the contribution of alternative splicing to eukaryote transcript diversity? I found that the previous "interchangeable model" of AS and gene duplication is a by-product of an existing relation between gene expression breadth, AS and gene family size. I also show that alternative splicing has played a key role in the expansion of transcript diversity and that this expansion is the best predictor reported to date of organisms complexity assayed as number of cell types. In addition, by comparing alternative splicing patterns in cancer and normal transcript libraries I found that cancer derived transcript libraries have increased levels of "noisy splicing".

1. Introduction

Alternative splicing (AS) is a post-transcriptional process in eukaryotic organisms by which multiple distinct transcripts are produced from a single gene (Graveley 2001). Recent studies using high-throughput sequencing technology have reported that up to 92%~94% of human multi-exon genes undergo AS (Pan et al. 2008; Wang et al. 2008), and AS has been proposed to be a major factor in expanding the regulatory and functional complexity, transcript diversity and organismal complexity of higher eukaryotes (Nilsen and Graveley 2010). Alternative splicing patterns are also frequently tissue/development-specific (Stamm et al. 2005; Wang et al. 2008), working independently of transcription regulation and providing an additional level of flexible control of gene expression (GE)(Artamonova and Gelfand 2007). Furthermore, both experimental and bioinformatics studies have shown that AS generates a variety of message RNA (mRNA) and protein products displaying distinct stability properties, subcellular localization and function (Stamm et al. 2005) thereby playing important roles in cell differentiation (Heinzen et al. 2008), sex differentiation (Blekhman et al. 2010; Hartmann et al. 2011) and development (Stamm et al. 2005), while aberrant AS may lead to cancer and disease (Venables et al. 2009; Watson and Watson 2010).

In the following sections I will briefly describe how alternative splicing was first discovered and the current understanding of this process and its regulation. I will then describe how AS is measured and what is known about its evolution and its role in physiological states and disease.

1.1 Alternative splicing

In 1977, Chow et al. reported that 5′ and 3'terminal sequences of several adenovirus 2 (Ad2) mRNAs varied, implying a new mechanism that the diversity of splicing patterns and the variety of recombined sequences generated during the synthesis of late Ad2 mRNAs, following this study, alternative splicing was also found in the gene encoding thyroid hormone calcitonin in mammalian cells (Berget et al. 1977; Chow et al. 1977; Alt et al. 1980; Early et al. 1980). Subsequent studies revealed that many other genes were also able to generate more than one transcript by cutting-out different sections from their coding regions (reviewed in (Graveley 2001; Artamonova and Gelfand 2007)).

Two decades after the discovery of alternative splicing, the first draft of the human genome sequence (Lander et al. 2001; Venter et al. 2001) was unveiled in February 2001 by two rival teams (the first an international consortium and the second, released by the private company CELERA). This draft of the human genome surprised academics as it showed our genome to contain ~23000 genes, only a fraction of the numbers of genes originally predicted (Crollius et al. 2000). In 2005, "Why the human genome has so few genes?" made it to the list of 25 top unanswered questions in science (http://www.sciencemag.org/site/feature/misc/webfeat/125th/), which drew much attention to alternative splicing given its potential to increase transcript diversity. With the development and subsequent constant improvement of whole genome transcription profiling and bioinformatics algorithms, the scale of occurrence of alternatively spliced transcripts began to become clear. Initial whole genome analyses suggested that 5%-30% percent of human genes were alternatively spliced. However, over the last ten years this number has been revised over and over with the latest estimates showing that up to 95% percent of human multi-exon genes produce more than one transcript through alternative splicing (reviewed in (Artamonova and Gelfand 2007)). The concept of one gene coding many proteins and prevalence of alternative splicing were gradually accepted as evidence mounted on the high percentage of AS incidence in human, mouse (Artamonova and Gelfand 2007) and other eukaryotes (Kim et al. 2007b).

1.2 Alternative splicing and its regulation

Splicing of precursor mRNA (pre-mRNA) is an essential step of gene expression in eukaryotes. Gene expression of protein-coding genes, the passage of information from the DNA gene sequence in a chromosome to the making of a protein- can be divided into two stages: transcription and translation. The process of transcription is controlled by transcription factors which bind in or near the promoter region, leading to the recruitment of RNA polymerases that copy the DNA nucleotide sequence into RNA. This RNA sequence contains a copy of the complete gene protein-encoding region (pre-mRNA) that contains both exons (sections which encode the sequence of amino acids in a protein) and introns (intervening segments which are removed or "spliced out" before the protein is produced). In the next stage, splicing is initiated by a complex of RNA-binding proteins known as the spliceosome, which catalyse the removal intron sequences of the pre-

mRNA. The recognition of intron-exon boundaries is facilitated by the detection of small sequence motifs called splicing sites. At the same time another process called polyadenylation (poly-A), which refers to the addition of a poly-adenine tail to an untranslated region located in the 3´-terminal end of the last exon, is initiated by a polyadenylating enzyme. Once the pre-mRNA has been processed by splicing and polyadenylation, the resulting mature mRNA is transported out of the nucleus into the cytoplasm where translation of mRNA occurs by the decoding machinery (the ribosomes) resulting in the assembling of a distinct polypeptide.

Splicing is tightly regulated by *cis* elements within exons and surrounding introns as well as *trans*-acting factors that bind to these *cis* element. Alternative splicing is an important mechanism of genetic control and its regulation is part of a complex network of regulatory events at different levels. Here I will describe different aspects of alternative splicing regulation.

In general, alternative splicing is thought to be controlled by RNA binding proteins that modulate the activity of the spliceosome which is composed of up of 5 small nuclear ribonucleoproteins (snRNPs) and more than 150 additional proteins. Besides the core components of the spliceosome, classic models of alternative splicing regulation also involve auxiliary splicing factors-proteins that are Serine/arginine-rich proteins (SR proteins) and heterogeneous nuclear ribonucleoproteins (hnRNP) proteins. SR proteins typically bind to exonic splicing enhancers (ESEs), where they interact with and recruit various components of the spliceosome such as U2AF to enhance 3´ splice site (3´ss) recognition and interact with U1-70K and recruit the U1 small nuclear RNP (snRNP) to process the adjacent 5´ splice site (Chen and Manley 2009; Graveley 2009). By contrast, hnRNP proteins tend to bind to exonic splicing silencers (ESSs) or intronic splicing silencers (ISSs) and repress splicing by other mechanisms (Chen and Manley 2009). For instance, hnRNP proteins prevent U2AF function by binding to the polypyrimidine tract. Also, hnRNP may bind to ISSs in the introns flanking an exon and looping the exon out of the pre-mRNA (Graveley 2009). Alternative splicing has also been shown to be regulated without the involvement of auxiliary splicing factors (Yu et al. 2008), suggesting the existence of additional non canonical mechanism of alternative splicing that are yet to be identified (Graveley 2009).

The regulation of alternative splicing is a multifactorial process also related to other types of events, such as, initiation of transcription from alternative promoters and alternative polyadenylation. Alternative mRNA isoforms may be subject to different

regulation by internal translation initiation sites, RNA editing, mRNA decay and microRNA binding (Hughes 2006). Recently, a direct role of histone modifications in alternative splicing has been reported, in which histone modification affect the splicing outcome by influencing the recruitment of splicing regulators via a chromatin-binding protein in a number of human genes (Luco et al. 2010). Additionally, non-coding RNAs also have emerged as key determinants of alternative splicing patterns (Luco and Misteli 2011) therefore revealing an additional layer in the regulation of alternative splicing. Recently, a machine-learning method based on the splicing code has been shown to predict tissue-specific expression with high efficiency (Barash et al. 2010), providing novel tools and methods of studying and predicting the outcome of alternative splicing.

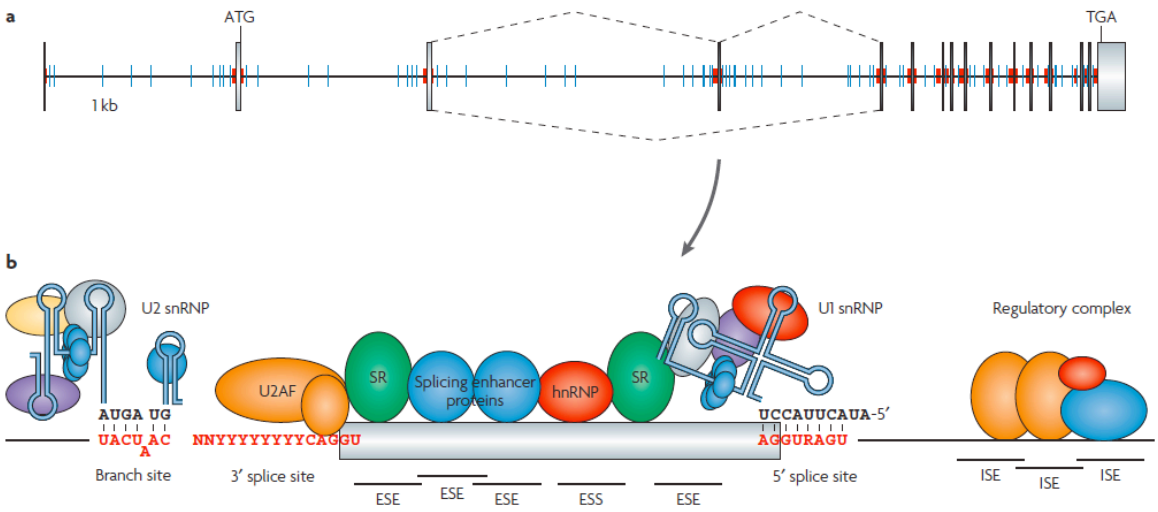


Figure 1: The splicing code. a. A exon/intron structure of pre-mRNA. b. A schematic of regulated splicing. Exons (boxes), introns (solid line). Splicing is regulated by cis-elements (ESE, ESS, ISS and ISE) and trans-acting splicing factors (SR proteins, hnRNP, and unknown factors). The 5' splice site (5'ss) and branch site serve as binding sites for the RNA components of U1 and U2 small nuclear ribonucleoprotein (snRNPs), respectively. This RNA base pairing determines the precise joining of exons at the recognition. This figure is adapted from (Wang and Cooper 2007b).

Depending on either the location of the exonic segments cut out or if introns are left in, AS events can be classified into five basic types (Figure 1). These five major modes of AS are: (1) Exon skipping (2) alternative donor site (5' ss), (3) alternative acceptor site (3'ss) (4) intron retention and (5) mutually exclusive exons (Ast 2004). In addition,

alternative initiation and alternative polyadenylation provide two other common mechanisms for generating various transcript isoforms. Moreover, different types of alternative splicing can occur in a combinatorial manner and one exon may be subject to more than one AS modes, for example, 5'ss and 3'ss at the same time. Differences in the relative prevalence of AS modes exist among different taxonomic lineages. For example, one comparative genomics study has shown that relative prevalence of the different types of AS varied from plants to metazoan, in which plants have higher level of intron retention while that of metazoan tend to be exon skipping (Kim et al. 2007a).

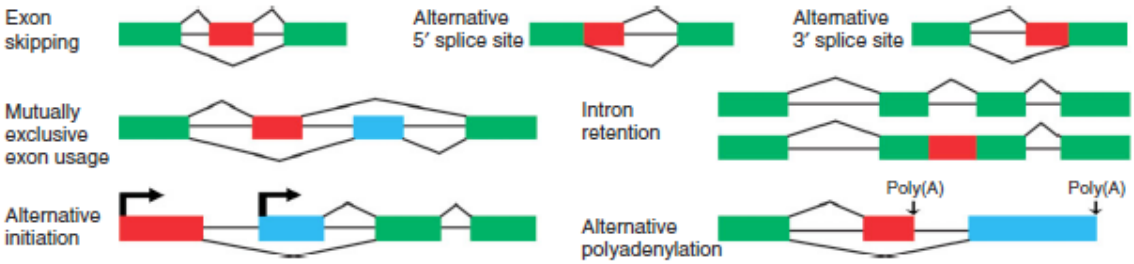


Figure 2: Different types of alternative splicing. The green boxes are constitutive exons and alternatively spliced regions in red. Introns are represented by straight lines between boxes and other lines indicate splicing activities. This figure is adapted from (Lu et al. 2011).

1.3 Measuring alternative splicing

Alternative splicing is difficult to estimate from genomic parameters (Barash et al. 2010). Regulatory motifs for alternative splicing are only now beginning to be uncovered and even the presence of these motifs may not guarantee that a gene is alternatively spliced (Barash et al. 2010). Accordingly, alternative splicing patterns have to be assessed from examining transcript data. For any alternatively spliced gene of interest, reverse transcription polymerase chain reaction (RT-PCR) can be conducted from a complementary DNA (cDNA) library in a specific tissue or development stage. A pair of forward and reverse PCR primers can be designed to target the flanking exons of a particular alternatively spliced exon. After the RT-PCR reaction, PCR products of different sizes corresponding to distinct mRNA isoforms can be separated and visualized

by electrophoresis. Combined with the use of radioactive or fluorescent labeling of PCR products (e.g. Quantitative PCR or Real-time PCR), this approach is highly sensitive and accurately reflects the splicing levels of individual exons. The main disadvantage of his method, however, is its low throughput level, making it suitable for detection or validation of a small set of genes rather than large-scale analyses. Over the last decade as high throughput transcriptome assessment technologies have improved, it has become possible to assess alternative splicing patterns on a genome wide scale. Three main sources of transcriptome data have been used to assess splicing patterns: expressed sequence tags (ESTs), splice-junction microarrays, and RNA-seq.

The first wave of genome-wide analysis of whole transcriptomes consisted in direct sequencing cDNA and ESTs carried out at large scale (Sayers et al. 2009). Alternative splicing events can be identified by aligning cDNA/EST sequences to the reference genome. In order to analyze alternative splicing in any RNA sample of interest in a global and quantitative manner, splice-junction microarrays and RNA-Seq were developed. Splicing microarrays target specific exons or exon-exon junctions with oligonucleotide probes. The fluorescent intensities of individual probes reflect the relative usage of alternatively splicing exons in different tissues and cell lines (Johnson et al. 2003). High-density splice-junction microarrays are a cost-effective way to assay previously known exons and alternative splicing events with low false positive rate. The drawback is that it requires prior knowledge of existing AS variants and gene structures. More importantly unlike RNA-Seq and EST, microarrays do not provide additional sequence information. In this thesis we have opted out from using data from splice-junction microarrays due to above-mentioned reasons apart from its limited data availability for different species (Blencowe 2006).

RNA-Seq has emerged as a powerful technology for transcriptome analysis due to its ability to produce millions of short sequence reads (Wang et al. 2009; Robertson et al. 2010; Martin and Wang 2011). Briefly, the protocol of RNA-Seq consists of the following steps (i) polyadenylate RNAs, (ii) convert into randomly sheared cDNA, (iii) select size of sheared fragments, (iv) amplify and ligate adapters to fragments and finally (v) sequence the fragments using next generation sequencing (NGS). Reads can be obtained from only one end of a fragment (single-end sequencing) or from both ends of a fragment (paired-end sequencing). RNA-Seq experiments provide in-depth information on the transcriptional landscape with high sensitivity and scale (Wang et al. 2009). The ever increasing accumulation of high throughput data will continue to provide ever richer

opportunities to investigate further aspects of AS such as low-frequency AS events as well as tissue-specific and/or development-specific AS events (Pan et al. 2008; Wang et al. 2008; Martin and Wang 2011; Ozsolak and Milos 2011).

ESTs are short (200–800 nucleotide bases in length), unedited, randomly selected single-pass sequence reads derived from cDNA libraries (Nagaraj et al. 2007). To obtain ESTs, mRNA sequences from expressed genes are first reverse transcribed to double-stranded complementary DNA (cDNA), which is further cloned to make cDNA libraries that represent a set of expressed mRNAs of the original cell or tissue. These cDNA clones are sequenced at random from both directions in a single-pass run of the polymerase. Currently, there are eight million ESTs for human (including 0.9 million sequences from cancer tissues) and about 71 million ESTs for around 2000 species (Boguski et al. 1993). I have chosen this data source for all the analyses presented in this thesis for two main reasons: First, ESTs still provide on average longer transcript segments than most publicly available RNA-seq data; splicing junction microarray data, on the other hand, do not provide any sequence information and available data is restricted to fewer species. Second, compared with both splice-junction microarray data and RNA-seq, EST data on public repositories are more abundant and comprehensive in terms of tissues and species covered so far. Regardless of the advantages of using EST data, we are fully aware of theis limitations given in that they are based on Sanger sequencing and are aggregated over a wide range of tissues, developmental states and diseases using widely different levels of sensitivity. We are expecting to update our AS database and test evolutionary hypotheses in a global and quantitative manner using RNA-seq in the near future (Hawkins et al. 2010; Martin and Wang 2011; Ozsolak and Milos 2011).

1.4 Alternative splicing in disease

Alternative splicing is essential for normal cellular functions. However, the *cis*- and *trans*-acting mutations, that disrupt the splicing code or the machinery required for splicing and its regulation, are known to cause disease (reviewed in (Brinkman 2004; Venables 2006; Wang and Cooper 2007b; Venables et al. 2009)). It has been estimated that 15-60 % of mutations that cause disease by affecting the splicing pattern of genes (Lopez-Bigas et al. 2005) (Wang and Cooper 2007b).

Specific mechanisms causing altered splicing have been described to fall within the following broad categories (Wang and Cooper 2007b; Jensen et al. 2009): (1)

disruption of the splicing code. The majority of disease-causing splicing mutations affect critical *cis* splicing regulatory signals such as mutations of the consensus splicing site sequences at particular exon-intron boundaries creating cryptic splice sites, or altering the secondary structure or regulatory complex-binding regions including splicing enhancer/silencer elements within exons or introns (Wang and Cooper 2007b). However, due to the complexity of splicing code, we are still a long way from accurately predicting whether a mutation or genetic variation will disrupt the splicing code and alter splicing patterns (Lu et al. 2011). (2) *Disruption of splicing machinery*. Disease-causing splicing mutations can act in *trans* as well. For example, frequent mutations of splicing pathway machinery were reported in myelodysplasia, providing the first evidence indicating that genetic alterations of the major splicing components could be involved in human pathogenesis also implicating a novel therapeutic possibility for myelodysplasia (Yoshida et al. 2011). (3) *RNA gain of function*, which is caused by trans-dominant effects on splicing regulation. For instance, repeated 3-10 nucleotides within coding regions expand beyond pathogenic thresholds cause microsatellite-expansion disorders (reviewed in Wang and Cooper (2007)). (4) *Disease-specific AS events*. Several studies have explored cancer related changes in alternative splicing patterns (reviewed in (Kalnina et al. 2005; Venables 2006; Skotheim and Nees 2007; Wang and Cooper 2007a)) resulting in the identification of an increasing number of cancer-specific AS events in a variety of cancer tissues, and these disease-specific AS changes have been proposed to play an important role in cancer development (Wang et al. 2003; Xu 2003; Hui et al. 2004; Kim et al. 2008a; He et al. 2009).

1.5 Prevalence of alternative splicing across eukaryotic genomes

Alternative splicing has been the subject of increased interest over the last few years with at least two whole genome studies in human having been published in quick succession (Pan et al. 2008; Wang et al. 2008). The number of species with reported AS events has continually increased over the years with instances of alternative splicing reported in plant and fungal species (Artamonova and Gelfand 2007; Kim et al. 2007a) demonstrating that AS appeared early in the evolution of eukaryotes. How prevalent alternative splicing is across different taxa and how alternative splicing patterns have changed and evolved through time, however, remains poorly understood. In fungi, AS is

thought to be rare due to the low number of introns in yeast (Ast 2004). In plants it has been estimated that around 20% of genes undergo AS (Wang and Brendel 2006). A better understanding of how alternative splicing has changed over time could provide a better understanding of how alternative splicing has impacted on transcript and therefore transcript diversity and phenotypic complexity (Nilsen and Graveley 2010). A few studies have attempted to compare alternative splicing prevalence between species with animals generally reported to have higher AS incidence than plants (Artamonova and Gelfand 2007) and vertebrates having a higher AS incidence than invertebrates (Kim et al. 2007a). The fact that alternative splicing identification rates is highly sensitive to transcript coverage (Kim et al. 2007a) (Brett et al. 2002) (Kan et al. 2002), makes it difficult to assess how alternative splicing prevalence varies across taxa (Nilsen and Graveley 2010). In order to systematically assess AS prevalence among different taxa, I created a database for 114 species of fully sequenced eukaryotes with at least 30000 ESTs available per species.

1.6 Structure of the thesis

Alternative splicing can modulate gene function, affect organismal phenotype and is widespread in eukaryotes but, how did it originate and evolve? Our understanding of its evolution is limited and comparative genomics of alternative splicing becomes critical in answering this question. In Chapter 2, I present the Eukaryote Comprehensive & Comparable Alternative Splicing Events Database (ECCASED), a new AS event database web resource. In addition to identifying as many AS events as possible, I also created an AS index which was comparable among genes within a species and between species with different transcript coverage. There are a number of databases which provide multispecies AS data (Kim et al. 2007b; Lee et al. 2007; Wang and Burge 2008; Bhasi et al. 2009; Koscielny et al. 2009b). However, these existing resources primarily focused on animal species given the poor coverage for protist, fungal and plant genomes. In addition none of these resources allow for comparative analyses of AS as they fail to correct for differential transcript coverage among genes within and among species. These biases result from the fact that for any two genes producing an equal number of AS isoforms, the more sequences there are available for one of them, the more likely it is to sample a more complete set of AS transcripts. Based in the ECCASED database, I was able to address a

number of fundamental questions on alternative splicing and its role in the eukaryotic transcriptome that are explored in consecutive Chapters.

Gene duplication and alternative splicing are two main contributors to transcript diversity. On the one hand, gene duplications create new duplicated genes and evolve functional divergence (Long et al. 2003), driving the evolution of developmental and morphological complexity in vertebrates (Dehal and Boore 2005). On the other hand, alternative splicing, as a common post-transcriptional process in eukaryotic organisms, has been proposed as a potential mechanism for the production of multiple transcripts encoding proteins with functional differences from a single gene (Graveley 2001; Nilsen and Graveley 2010). In Chapter 3, I explore how alternative splicing relates to gene duplication by testing the relationship between gene family sizes and AS levels and different AS levels in 17 species from plants to mammalians. In Chapter 4, I focus on how alternative splicing has contributed to proteome expansion and phenotypic complexity over 1400 million years.

Given the high number of AS events unique to cancer transcriptomes, cancer-specific transcripts have been proposed to play a key role in cancer physiology (Skotheim and Nees 2007; He et al. 2009). A number of studies however have shown that a significant proportion of AS transcripts are likely to be the result of alternative splicing noise and are unlikely to produce a functional protein (Green et al. 2003; Lewis et al. 2003; Zhang et al. 2009; Pickrell et al. 2010). Whether splicing events specific to cancer genomes are likely to contribute to cancer onset or cancer maintenance have not been explored in any depth. In chapter 5, I explore the likelihood of cancer-specific AS events being functional in the cancer transcriptome.

2 ECCASED: Eukaryotic Comprehensive and Comparable Alternative Splicing Events Database for 114 eukaryotic species

2.1 Introduction

Alternative splicing (AS) is a common post-transcriptional process in eukaryotic organisms by which multiple distinct functional transcripts are produced from a single gene by selectively cutting out segments from RNA transcripts (Graveley 2001). AS plays a key role in the regulation of transcript diversity in eukaryotic genomes with over 92%~94% of human multi-exon genes now known to undergo AS (Pan et al. 2008; Wang et al. 2008). The dramatic increase of fully sequenced genomes and transcript data availability (Boguski et al. 1993) opens up the opportunity to assess AS patterns for genes within and across species providing insights into the evolution of AS (Artamonova and Gelfand 2007), transcript diversity and the evolution of complexity (Nilsen and Graveley 2010).

There are a number of databases that provide AS data for multiple species (Kim et al. 2007b; Lee et al. 2007; Bhasi et al. 2009; Koscielny et al. 2009a). However, these existing resources are primarily focused on animal species and have poor coverage for protist, fungal and plant genomes. Most importantly, none of these resources take into account the effects of differential transcript coverage across genes within and between species which greatly influences AS detection rates (Brett et al. 2002; Kan et al. 2002; Kim et al. 2007a; Nilsen and Graveley 2010).

Here we present a Eukaryote Comprehensive & Comparable Alternative Splicing Events Database (ECCASED) based on the analysis of 39,620,288 EST transcripts available at dbEST (Boguski et al. 1993) for 114 eukaryotic genomes, including protists (20), plants (20), fungi (23) and metazoans (51, including 18 vertebrates). Using a uniform pipeline across all species, ECCASED is the most comprehensive database available to date in terms of the number of species covered and transcripts analyzed. Unlike other resources, ECCASED also provides a comparable AS index based on

random sampling allowing direct comparisons of AS in genes within and between species.

2.2 Materials and methods

Genome sequences, gene annotations and transcript data were downloaded from several publicly available genomic data sources during May 2011 (for a full list of data sources per species see Supplementary Table 1 at http://bio.bdfield.com/eccased/downloads.php).

To identify AS events, the following analysis pipeline was used (see Figure 1):

(i) *EST to gene matching*. Each EST was aligned against the corresponding genome sequence for its species using Genomic Mapping and Alignment Program (GMAP) software (Wu and Watanabe 2005). ESTs were required to align to the genome sequence with at least 95% identity and 95% coverage of its length; those which failed this requirement were removed from further analyses. Each EST was associated to their best hit according to identity and coverage as provided by the GMAP software (Wu and Watanabe 2005). ESTs matching to regions with no annotated genes were discarded from further analyses. Any ESTs with a best hit to a region with an annotated gene (from start of first to end of last exon) was assigned to that gene. All annotated overlapping genes, including any instances of nesting and their matching ESTs were also removed from further analyses. Genes with no matching transcripts were removed from the AS identification pipeline but all other gene annotations and functional classifications were retained for database construction. All ESTs from cancer-derived EST libraries were removed from further analyses.

(ii) *Template building*. To obtain an exon template as complete as possible (as well as overcoming the fact that some invertebrates do not have full transcripts sequenced) all available ESTs for a given gene were overlaid onto its genomic sequence (Figure S1A and S1B). First the longest partial or full transcript available forms the base of the template. All other mRNAs and ESTs are then aligned to the genomic sequence and boundaries with the previously included transcripts are revised to extend exons or include new ones. If a transcript only encompasses a single exon then it will be discarded. This allows identifying and discarding any single exon nesting genes that have not been previously annotated.

Over 35000 exons across all species which were supported by fewer than 5 percent of ESTs available for any given gene were removed from further analyses to avoid inclusion of exons resulting from splicing errors.

(iii) *Detection of AS events.* Exon boundaries for each EST (obtained from GMAP alignments) were compared to its corresponding gene template (Figure S2) to identify AS events. AS events were classified into 8 different types (classification adapted from (Malko et al. 2006); Figure S2). For the purpose of counting AS events per gene, any AS events with coordinates differing by less than 15bp were considered as one. This is because the algorithm used by GMAP depends on an 8-mers finding clusters algorithm (Wu and Watanabe 2005). Coordinates of all AS events identified in each EST are available as bulk downloads per species.

(iv) Calculating comparable AS indexes. In addition to AS event counts from all ESTs available per gene, we also obtained a comparable index which avoids biases due to differential transcript coverage. For this, the average number of AS events in 100 randomly selected samples of 10 ESTs was calculated for all genes with over 10 associated ESTs (Kim et al. 2007a). It is important to note that the comparable index is not intended to reflect the number of AS events per gene but instead the number of AS events found per 10 ESTs.

(v) Identification of AS isoforms. To identify AS isoforms, ESTs with at least one AS event were first sorted according to the number of AS event they contain. Then ESTs containing identical or similar AS events were classed as redundant and excluded from the analysis. The number of remaining ESTs was taken as estimate of AS isoforms produced per gene. Result tables contain coordinates and EST support numbers for each AS event. AS event annotation per EST are provided in the download files.

Additional data including Gene ontology terms, gene description and homologous relationships were retrieved from BioMart (Haider et al. 2009) for 71 species. For the remaining 43 species not currently supported by Biomart orthology relationships were assessed using the InParanoid software (Berglund et al. 2008). EST expression annotations were retrieved from EST library information contained in the dbEST bulk file (Boguski et al. 1993).

The ECCASED database is in accordance with the format of BioDBcore (http://biocurator.org/biodbcore.shtml).

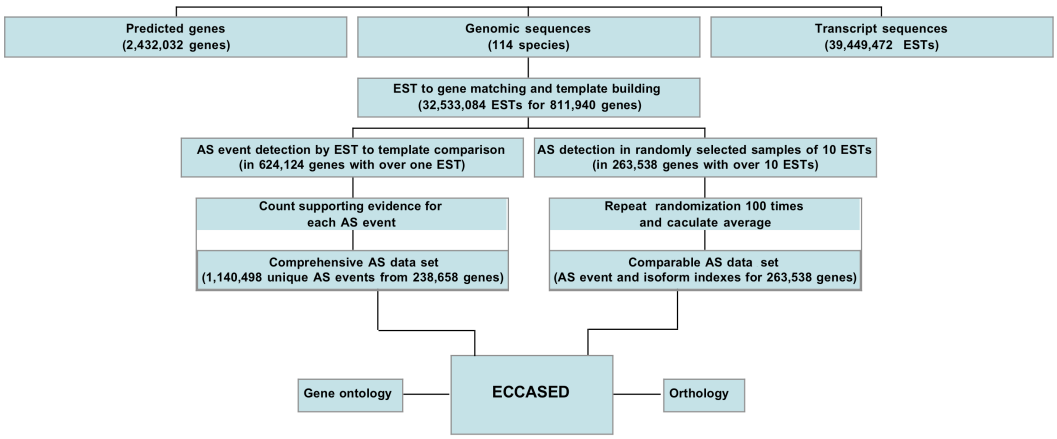


Figure 1. Overview of the ECCASED database building pipeline. Starting from genomic and EST transcript sequences and gene coordinates (see supplementary Table 1 for sources), individual ESTs were matched to specific genes. After producing a full template of intron/exon coordinates, AS events were detected in either all ESTs available per gene or within 100 samples of 10 transcripts. Unique AS events and isoforms were identified by removing redundant transcripts containing AS coordinates differing by less than 15bp from each other. Gene ontology annotations and gene orthology relationships are integrated to AS comprehensive and comparable data (see methods for further details).

2.3 Results

2.3.1 The ECCASED database

In total, 39,449,472 ESTs were analyzed to identify instances of AS events corresponding to 811,940 genes out of a total of 2,432,032 annotated genes in 114 species analyzed including 23 fungi, 20 protists, 20 plants and 51 metazoans including 18 vertebrates (see methods and Figure 1). A total of 1,140,498 unique AS events corresponding to 238,658 genes across all species were identified across all species (see methods and Supplementary Table 3). Importantly, while other EST based AS databases identify AS events in 40-60% of human genes (Kim et al. 2007b; Lee et al. 2007; Bhasi et al. 2009), we found that 97% of human multi-exon genes are alternatively spliced with >10 ESTs per gene, which is in line with recently reported high-throughput sequencing based studies (Pan et al. 2008; Wang et al. 2008). In other vertebrates, on average 89.4% were classed as undergoing AS, the highest rate reported to date (Kim et al. 2007b; Lee et al. 2007; Bhasi et al. 2009).

Table 1. Statistics summary for ECCASED

Eukaryotic groups	Species	Annotated genes	Genes with EST	All AS genes	Genes (>10 ESTs)	AS genes (>10 ESTs)	AS prevalence (>10 ESTs) %
Fungi	23	258042	84982	7312	11080	3979	35.9
Protists	20	399957	93087	6370	13669	3143	23.0
Plants	20	709185	248710	72664	156238	52757	33.8
Insects	15	258552	71558	14620	20947	10429	49.8
Invertebrates (other)	18	406564	101394	19874	24252	12188	50.3
Vertebrates	18	399732	212209	117818	101212	90505	89.4
Total	114	2432032	811940	238658	327398	173001	52.8

The ECCASED database represents the first assessment of alternative splicing patterns for approximately 70% of species analysed. The analysis of 20 plant genomes showed that more than 33.8% of genes undergo alternative splicing (Table 1 and Supplementary Table 3) compared to 20% in previous estimates based on a handful of species and using fewer transcripts (Wang et al. 2008). Interestingly, in fungal species where AS was previously thought to be rare (Ast 2004), we found evidence for AS events in 35.9% of genes with more than 10 ESTs (Table 1 and Supplementary Table 3). Protists were the group with the lowest overall AS incidence with just over a fifth of genes (23.0%) found to undergo alternative splicing (Table 1 and Supplementary Table 3). To our best knowledge, ECCASED is to date both the most comprehensive and the largest alternative splicing database of eukaryotic genomes.

2.3.2 Comparable AS estimates

There is a strong dependence of AS detection on transcript coverage across genes within and between species which has hampered the comparative analyses of AS patterns (Brett et al. 2002; Kim et al. 2007a; Nilsen and Graveley 2010). For example, while on average there are over 150 transcripts for every mouse gene, for its closely related species the rat there are, on average, little over 30 transcripts per gene. As a result, on average 3.03 isoforms per gene are identified in mouse while in the rat the average number of AS isoforms detected is 1.70 (Supplementary Table 3). By using a random transcript sampling method to obtain comparable AS estimates (see methods), we minimize the

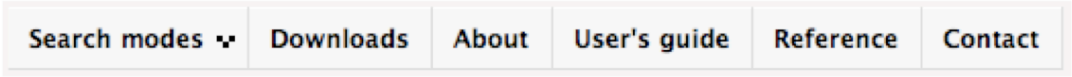
dependence between AS and the ESTs coverage within and across species, resulting in AS indexes more similar for closely related species regardless of differences in transcript coverage. In the case of mouse and rat, using the random sampling protocol, we obtained an average of 1.47 and 1.41 in each sample of 10 ESTs respectively. This comparable AS index is not a calculation of the total number of AS events in a gene but instead is an index of AS events per ten transcripts to allow a direct comparison of genes within a genome and for genes in different species with differing transcript coverage (Kim et al. 2007a). Based on this comparable AS data, we also calculated the number of genes with at least one alternative splicing event in every ten ESTs. This AS prevalence index is likely to be an underestimation of AS prevalence as most genes produce many more transcripts however it allows for a direct comparison of relative differences in AS prevalence among species groups. We found that 90% of human genes have at least one on ten EST with alternative splicing evidence. Using the same comparative threshold we found that 21.3 of fungi, 13.0% of protist, 22.1% of plant, 33.5% of invertebrates and 79.8 of vertebrate genes have at least one EST in ten with AS evidence. To further facilitate the comparative analyses of the AS data in the ECCASED database we integrated gene homologous relationships among 71 species from BioMart (Haider et al. 2009). Homologous relationships for the remaining 43 species were generated using InParanoid software (Berglund et al. 2008) constituting, for many species, the first assessment of homology relationships.

2.3.3 Web interface and output

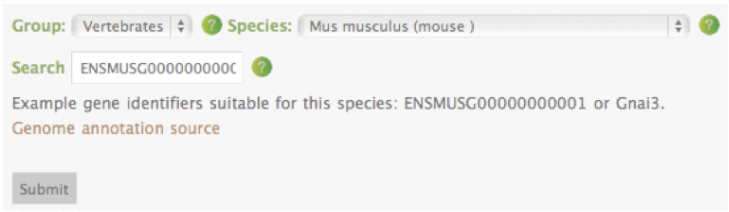
The ECCASED database can be consulted through a user-friendly web interface which provides: (i) AS summary statistics including the number of identified AS events and isoforms as well as AS comparable indexes based on random sampling analyses. Available transcript number is also provided (Figure 2A); (ii) exon information: genomic coordinates per exon as assembled de novo from the available ESTs, with supporting transcript number and AS event and ratio inside this exon. (iii) AS event listing with AS event genomic coordinates, type of AS event, transcript number evidence, and a representative transcript ID for each isoform with its expression information (a full list of mapping coordinates for AS events per EST is available for download); (iv) graphic genomic view showing gene template as well as representative transcript coordinates containing all AS events found per gene (Figure 2B); (v) homologous relationships along with AS statistics for each homologous gene (Figure 2C); (iv) gene ontology associated

terms for the gene. In addition, for users intending to use the database for genome wide studies, (v) bulk downloads of AS data per species or with genes divided according to GO terms are available.

A



B



C AS statistics

Species name	TaxID	Gene ID	Entrez ID	Symbol	Transcript num	Avg exon num per transcript	Total AS events	Total AS isoforms	AS events per 10 transcripts	AS isoforms per 10 transcripts
Mus musculus	10090	ENSMUSG00000000001	14679	Gnai3	384	4	4	4	0.97	0.97

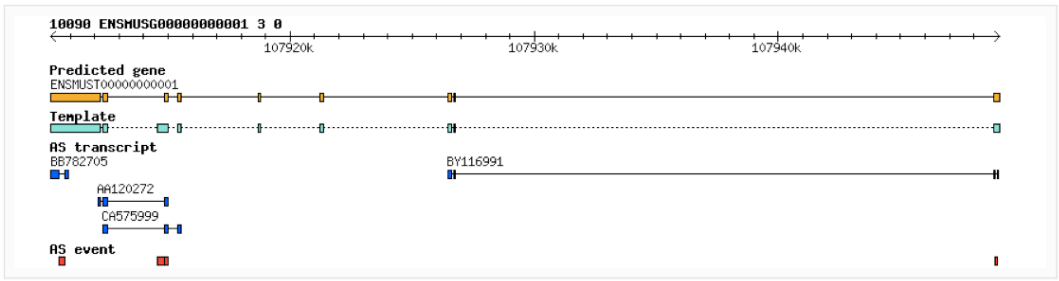
D Exon detail

Exon ID	TaxID	GeneID	Exon num	Exon coordinates	Exon evidence	AS ID	AS coordinates	AS event type	Ratio
20807158	10090	ENSMUSG00000000001	9	3:107910198-107912234	98	714827	107910543-107910773	I	0.020
20807159	10090	ENSMUSG00000000001	9	3:107912318-107912533	114	0			0.000

E AS events

TaxID	Gene ID	Entrez gene id	Gene symbol/name	AS coordinates	AS event type	AS event evidence	Representative transcript ID	Cell type	Cell line	Developmental stage	Sex	Tissue
10090	ENSMUSG00000000001	14679	Gnai3	3:107914853-107915006	3S5S	58	CA575999			Age approx.10 weeks old		
10090	ENSMUSG00000000001	14679	Gnai3	3:107910543-107910773	I	1	BB782705		CRL-2116 JC	1.5 years	female	

F Genomic view



G Homologous Genes

TaxID	Gene ID	Ortholog Species Name	Ortholog TaxID	Ortholog Ensembl ID	Ortholog Entrez ID	Ortholog Symbol	Ortholog Transcript num	Ortholog Avg exon num per transcript	Ortholog Total AS events	Ortholog Total AS isoforms
10090	ENSMUSG00000000001	Acyrthosiphon pisum	7029	ACYPI004308	100163208	XP_001948628.1	9	1.2	2	1
10090	ENSMUSG00000000001	Acyrthosiphon pisum	7029	ACYPI009431	100168757	NP_001129194.1	12	2.1	1	1
10090	ENSMUSG00000000001	Aedes aegypti	7159	AAEL008641			19	3	1	1

H Gene ontology

Ensembl Gene ID	Entrez gene id	Gene symbol/name	GO term	GO name
ENSMUSG00000000001	14679	Gnai3	GO:0007186	G-protein coupled receptor protein signaling pathway
ENSMUSG00000000001	14679	Gnai3	GO:0007264	small GTPase mediated signal transduction
ENSMUSG00000000001	14679	Gnai3	GO:0006906	vesicle fusion

Figure 2. Screenshot of the ECCASED page. From the top bar (A), Search modes, namely, gene identifiers, multiple genes, gene ontology and keywords can be selected. Links to downloads, readme files and contact instructions are also found here. Group and species restrictions can be applied by using drop down menus (B). Query output contains several tables: AS statistics (C) including comparable AS estimates; Exon information (D) which includes exon annotations with transcript support per exon and AS information; AS events (E) present coordinates, representative ESTs and its expression; genomic view (F) showing the gene template, representative ESTs for each AS isoform and alternatively spliced regions; AS summary statistics for homologous genes (G) and gene ontology annotations (H).

2.3.4 Database mining and tool

The ECCASED database can be queried in four main ways:

(A) Gene identifiers. Ensembl gene IDs are used as primary identifiers for genes of 71 species with various source specific identifiers used for the remaining 43 (see Supplementary Table 1 for a full list of gene annotation source data, gene ID type and format). Gene symbol/gene name and Entrez IDs can also be used in most species. Users can search for individual or multiple genes and restrict the searches to individual species, a predefined list or by group (Fungi, Protists, Plants, Metazoans (non-vertebrates) and Vertebrates) (Figure 2).

(B) Keywords (e.g. troponin) from gene descriptions. Advanced queries are also possible, for example, the query "+troponin –similar" will return every gene containing the keyword troponin but which does not contain the term similar on its gene description. By using double quotes, users can also specify the certain combination of word and character or number e.g. "troponin I".

(C) Gene ontology IDs and terms. Searches by gene ontology keywords (from GO term or keyword in GO description) are also supported. As the some GO terms are associated with a high number of genes, GO searches are restricted to a single species at a time. AS statistics of genes in the queried GO category will be shown, and can be downloaded for further AS comparison among GO categories.

(D) Bulk data download with AS events, AS statistics and gene template coordinates per species or by GO term are available. A more comprehensive download with mapping coordinates for AS events in each EST is also available. Users may request the AS identification software to analyze their own sequences.

2.4 Discussion

ECCASED provides multiple-species alternative splicing analysis based on public data in May 2011. As more genomic data becomes available, we plan to update ESTs analyzed, exon/intron gene annotations, GO ontology as well as the functional annotation of AS event/isoform. We plan to provide functional annotations of predicted Open Reading Frames (ORFs) as well as putative functional domains within constitutive and AS regions. With the recent advent of next-generation sequencing technologies, we are aware of the importance of using high-throughput sequences from RNA-seq in our analysis. We plan to integrate short read data with EST data to be able to provide quantitative expression abundance data on a per AS-isoform basis. An expression pattern based search will also be included

The ECCASED database covers 114 species with sequenced genomes making it the most comprehensive database to date. Unlike existing AS data resources, in addition to the listing of all identified AS events, the ECCASED database also provides a comparable AS index per gene avoiding the strong biases caused by differential transcript coverage in AS detection rates allowing direct contrast between genes within and across species. ECCASED data can be accessed in several ways from single gene analyses to bulk downloads to maximise appeal for a wide range of users and we expect this resource to facilitate future studies of alternative splicing patterns.

2.5 Supplementary Materials

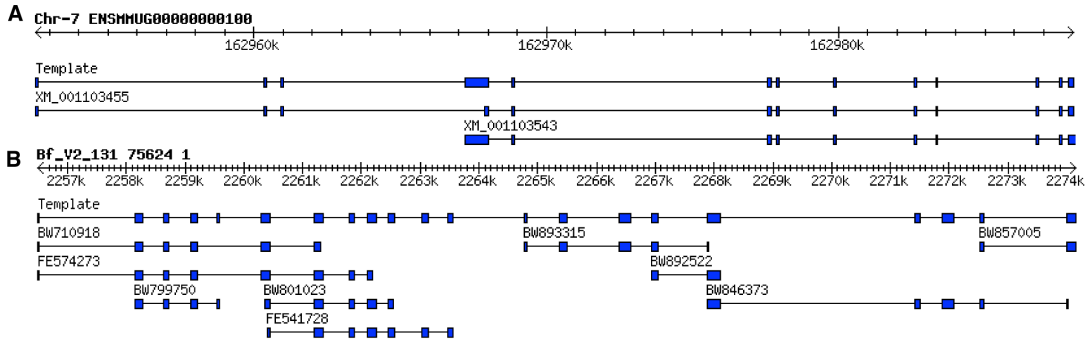


Figure S1. Exon template building. Panels A and B show gene templates for ENSMMUG00000000100 and 75624 from mouse and amphioxus, respectively. In both cases, the top line represents the chromosome region where the gene is located. The second line represents the resulting template from alignments of all available EST sequences. Note that exons supported by fewer than 5% ESTs for any given gene were disregarded as possible splicing errors. The following lines show the mRNAs and ESTs used to build the template. In the first example, by comparing two transcripts an exon segment was recovered (A). In the second example, a gene template was constructed from ESTs alone as no full length mRNA was available (B).

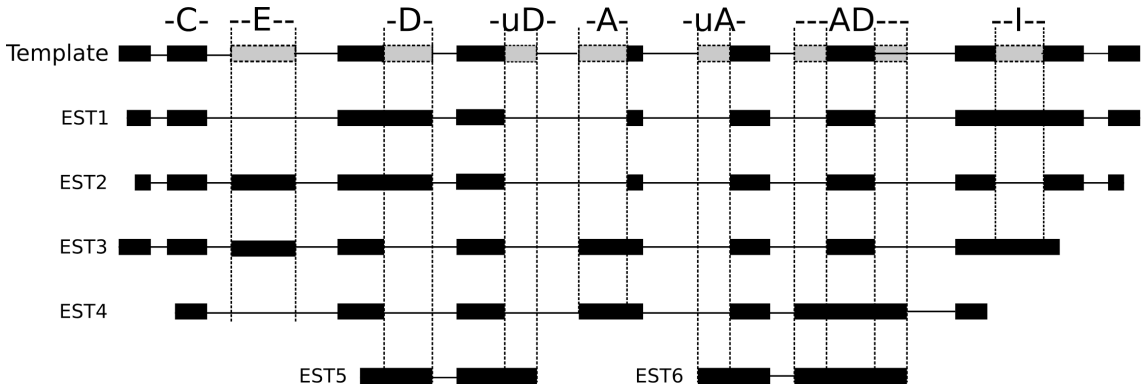


Figure S2. Identification of AS events. For each gene, AS events were identified by comparing individual transcripts against the template. In the diagram, constitutive exons are shown in black and AS events are shown in gray. Eight types of AS events were identified: Constitutive exon (C); Exon skipping (E); alternative 5' donor site (D); uncertain edge 5' donor site (uD); alternative 3' acceptor site (A); uncertain edge 3' acceptor site (uA); 3' acceptor site and 5' donor site (AD) and intron retention (I) (Malko et al. 2006). Splicing events with coordinates within 15bp of each other were considered to correspond to the same AS event.

3 Gene expression breadth explains the relationship between alternative splicing and gene duplication

3.1 Introduction

Both alternative splicing (AS) and gene duplication (GD) have been proposed to play important roles in the evolution of novel functions (Graveley 2001; Long et al. 2003; Dehal and Boore 2005; Nilsen and Graveley 2010). AS was once thought to be restricted to a small proportion of genes but recent studies have revealed that AS is prevalent in many eukaryote genomes (Artamonova and Gelfand 2007; Kim et al. 2007a) and in human 92%~94% of multi-exon genes undergo AS (Pan et al. 2008; Wang et al. 2008).

A number of studies exploring how these two mechanisms relate to each other have consistently found a negative correlation between gene family size (GFS) and average AS events detected per gene in human, mouse (Kopelman et al. 2005; Su et al. 2006; Jin et al. 2008) and worm (Hughes and Friedman 2008; Irimia et al. 2008). Singletons have been found to go against the inverse correlation by having lower alternative splicing than those gene families of two members (Jin et al. 2008; Roux and Robinson-Rechavi 2011) thus the negative correlation between AS and GFS may only apply to multi gene families. The correlation between AS and GFS has been explained as the result of the steady increase in alternative splicing per gene along time (Roux and Robinson-Rechavi 2011). The small number of species tested and diversity of datasets and methodologies used (Table S1), does not allow a comparison of findings for different species. Here, we systematically investigate the relationship between AS and GFS in 17 species from different taxa and assess the role of expression measures in driving this co-variance.

3.2 Materials and methods

3.2.1 Datasets

Genome sequences and predicted genes were downloaded from databases shown in Table S2, and EST sequences were downloaded from UniGene (Sayers et al. 2010)(ftp://ftp.ncbi.nih.gov/repository/UniGene/).

3.2.2 Identification of paralogs and orthologs

Using the method from a previous study (Jin et al. 2008), we assembled gene families from sea urchin and amphioxus according to the alignment of protein sequences. For predicted genes from Ensembl, we extracted paralogs of each Ensembl gene family ID from BioMart (Haider et al. 2009) and calculated the gene family size by adding the number of the paralogs and the gene itself. To identify orthologous relationship for Ensembl genes, the orthologs were retrieved from BioMart (Haider et al. 2009). We defined 3879 gene families with one or more genes that were present in at least three invertebrate and three vertebrate species. To assess whether families associated with different functional categories tended to have a decreased, stayed stable, or increased in transcript diversity, we first assigned non-redundant GO slims (Harris et al. 2004)from human to each gene family and then obtained linear regressions between transcript diversity against time of divergence from human lineage. Gene families then were assigned to one of 4 blocks according to the slopes from linear regressions per family and divergence time (see methods): ‘decreased’ (slope < 0.0000); ‘stable’ (0.0000 =< slope < 0.0015); ‘increased’ (0.0015 =< slope <= 0.0025); ‘highly increased’ (slope > 0.0025).

3.2.3 Identification of alternative splice events

To estimate AS events in different organisms, a novel procedure was applied as follows:

(1) *Mapping predicted genes and ESTs to Genome and grouping ESTs for each gene.* Overlapping and nested genes were identified and removed from further analyses. GMAP (Wu and Watanabe 2005) was used to align full transcripts and high quality ESTs to their corresponding predicted genes. Genes with no matching transcript were removed.

(2) *Template building*. To obtain a gene template as complete as possible (as well as overcoming the fact that some invertebrates do not have full transcripts sequenced) full transcripts and ESTs were overlaid onto the genomic sequence. This was done as follows: First the longest partial or full transcript available forms the base of the template. All other mRNAs and ESTs are then aligned to the genomic sequence and boundaries with the previously included transcripts are revised to extend exons or include new ones. If a transcript only encompasses a single exon then it will be discarded. This allows

identifying and discarding any single exon nesting genes that have not been previously annotated.

(3) *Detecting AS events and AS isoforms.* We developed an algorithm to compare the exon boundaries of any transcript to its corresponding template. To identify AS isoforms, transcripts were first sorted according to the number of AS events they contain. Then transcripts containing identical or similar AS events were classed as redundant. In addition to the listing of all identified unique AS events, we also generated a comparable AS index that minimizes the effects of differential transcript coverage. For this, one hundred samples of 10 randomly selected transcripts were obtained, for genes with at least 11 associated ESTs, in every species (Kim et al. 2007a). AS event and isoform number were then calculated as described above in each sample and results were averaged across all 100 samples per gene.

3.2.4 Gene expression data

Gene expression data for UniGene were downloaded from NCBI ftp (Sayers et al. 2010). We assign each EST to library according to its tissue, development state and whether it is from a cancer source or normal tissue. In this study, all libraries from cancer sources were excluded. In order to compare expression across species, we grouped library from different tissues into 10 common organ levels (http://bodymap.jp/organ_tissue_rule) that are comparable between species according to BodyMap-Xs (Ogasawara et al. 2006). In order to minimize the bias caused by different abundance (the number of ESTs) of the grouped libraries in the 10 common organs, we employed a random sampling to reconstruct the library, in which we randomly selected 10000 ESTs for 100 repetition (one million ESTs) from the pooled libraries of an organs, then counted how many times each gene present in this random sample of one million ESTs, which were used as a proxy of expression for each gene in different organs.

For independent expression data, we downloaded gene expression data for 10 species from BodyMap-Xs (Ogasawara et al. 2006). According to the corresponding table between Ensembl gene and UniGene, we joined UniGene expression data to Ensembl gene and remove any UniGene corresponding to more than one Ensembl gene. Microarray data in human and mouse (Su et al. 2004) and serial analysis of gene expression (SAGE) tag sequences in human were used (http://cgap.nci.nih.gov/SAGE).

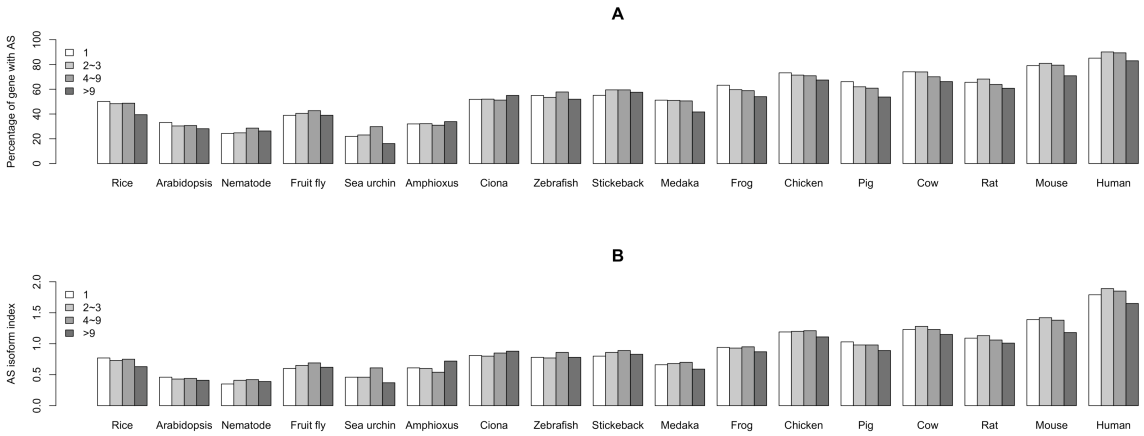


Figure 1: The comparison between the gene family size and the proportion of genes with AS and AS isoform index. Histogram bars indicate the fraction of genes containing more than one AS isoform (A) and the AS isoform index (B), classified as singletons (1 member) and small (2~3 members), medium (4~9 members) and large (>9 members) gene families.

3.3 Results

3.3.1 No universal negative correlation between gene family size and AS

To investigate the relationship between AS events and gene duplication, gene family annotations were obtained for 17 species (for species list see Table S2) and AS events per gene were identified using publicly available ESTs and mRNAs (see methods). We then assessed the relationship between AS and GFS. We found significant inverse correlations for only seven of the 17 species analysed ($P < 0.0029$ after Bonferroni correction). Singletons have been previously found to have lower alternative splicing than multi-gene families (Jin et al. 2008; Roux and Robinson-Rechavi 2011) and have been suggested to have different evolutionary paths compared to multi-gene families (Jin et al. 2008) and to have a slower gain of alternative splicing events (Roux and Robinson-Rechavi 2011). Although we did not find the lower AS levels for singletons to be a consistent pattern (Figure 1 and Table S3), we reassessed the relationship between both variables after removing singletons. A significant negative correlation coefficient was obtained for a total of six species ($P < 0.0029$ after Bonferroni correction, see Table 1).

Table 1. Correlation test between gene family size and AS occurrence in 17 species.

Organism	All Genes	*P*	*R*	Duplicates	*P*	*R*
Oryza sativa	13616	4.87E-23	-0.0846	11008	2.08E-22	-0.0926
Arabidopsis thaliana	12440	0.0048	-0.0253	10151	0.0548	-0.0191
Caenorhabditis elegans	4519	0.0097	0.0385	2568	0.4647	-0.0144
Drosophila melanogaster	5412	0.1038	0.0221	3012	0.0424	-0.0370
Strongylocentrotus purpuratus	1039	0.1582	0.0438	493	0.9357	0.0036
Branchiostoma floridae	1380	0.8334	0.0057	577	0.8991	0.0053
Ciona intestinalis	5117	0.3056	0.0143	2925	0.0062	0.0506
Danio rerio	8645	0.1074	0.0173	6385	0.8978	0.0016
Gasterosteus aculeatus	1742	0.2690	0.0265	1290	0.7611	-0.0085
Oryzias latipes	2817	0.0190	-0.0442	1817	0.0003	-0.0854
Xenopus tropicalis	5265	0.0023	-0.0419	3560	0.0067	-0.0455
Gallus gallus	5499	0.1697	-0.0185	3438	0.4509	-0.0129
Sus scrofa	5281	1.30E-06	-0.0665	3345	0.0054	-0.0481
Bos taurus	8420	1.08E-06	-0.0531	5794	1.99E-06	-0.0624
Rattus norvegicus	7665	0.0013	-0.0368	5413	2.45E-07	-0.0701
Mus musculus	13417	8.66E-22	-0.0827	9502	1.09E-37	-0.1311
Homo sapiens	13290	1.97E-05	-0.0370	9298	5.19E-29	-0.1156

A consistent downward trend in AS for larger gene families was not observed either when visually inspecting the data after dividing all gene families into four groups according to Kopelman et al. (2005): singletons; and families with 2~3, 4~9 and families with >9 members or when considering all families or multi-gene families only (Figure 1).

Given the large differences in transcript coverage among genes and between species and the fact that alternative splicing detection is highly dependent on transcript coverage and could thus influence the relationship between gene family size and alternative splicing, we calculated a comparable index of alternative splicing using random samples of ten transcripts (see methods). Using this normalised AS data, we found similar results to those using the non-corrected AS values with 3 species for all genes and 7 species for duplicates found to have a significant negative correlation for AS and GFS ($P < 0.0029$ after Bonferroni correction; Supplementary Figure 1; Table S4). Our results show that the inverse relationship between AS and GFS is not universal. Furthermore, the previously reported lower AS values for singletons do not hold true for seven of 17 species analysed. It is worth noting that consistent with previous reports (Kopelman et al. 2005; Su et al. 2006; Jin et al. 2008), significant negative correlations

were observed for human and mouse but was not recovered in the nematode (Hughes and Friedman 2008; Irimia et al. 2008).

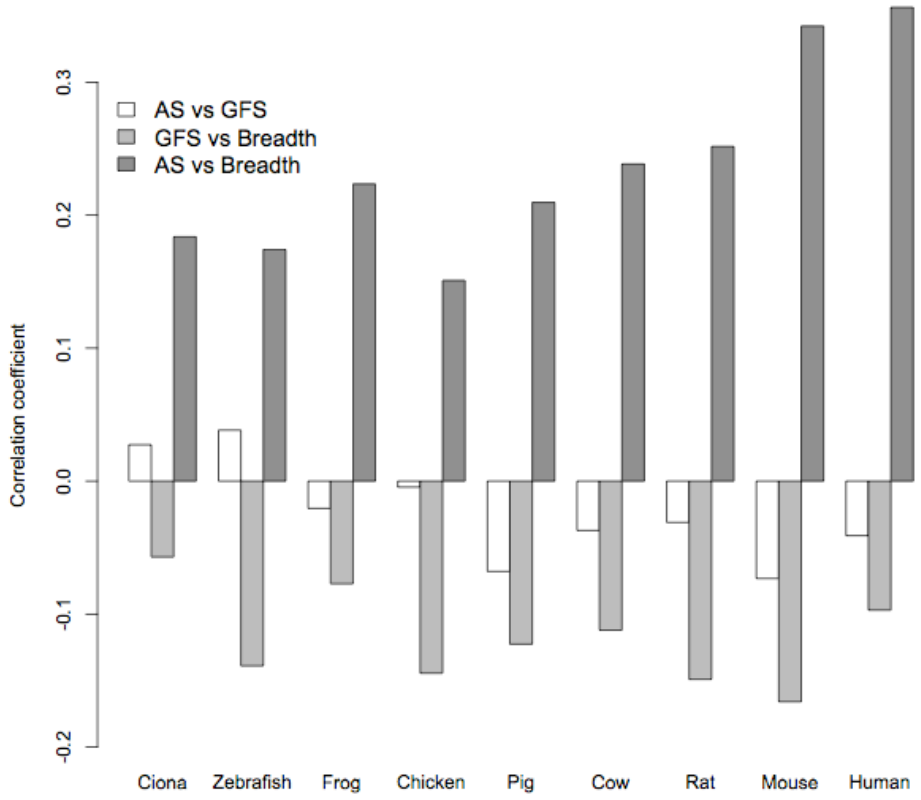


Figure 2: Comparsion of the relationship beteen AS, GFS and breadth. The histogram bars indicate the correlation coefficient of AS versus GFS, GFS versus Breadth, AS versus breadth in nine species, respectively.

3.3.2 The relationship between alternative splicing, gene family size and gene expression

It has been reported that width of gene expression (number of tissues where a gene is expressed) is linked to the gain of transcript isoforms (Wegmann et al. 2008). In addition, there is a general trend for duplicated genes to become more specialized in their expression patterns, with decreased breadth and increased specificity of expression per gene as gene family size increases (Huminiecki and Wolfe 2004; Freilich et al. 2006; Farre and Alba 2010). It is therefore possible that the observed weak but significant relationships between GFS and AS observed in some species could be the by-product of

the relationship of both variables with breadth of expression. Thus, we first assessed the relationship between AS (normalised) and GDF with three measures of gene expression: peak, mean and breadth in nine of the 17 species studied for which expression data was available (see methods). We found a consistent positive correlation between the breadth of expression with AS in all species, while a negative correlation between the expression breadth and GFS (Figure 2). Significant correlations for AS and GFS were also observed in some but not all species (Figure 2). Given that all three expression measures highly co-vary with one another (Lercher et al. 2002; Urrutia and Hurst 2003), we assessed whether all three were independently related to both AS and GFS. To do this, we performed forward stepwise tests where the contributions of the three measures of expression in predicting AS and GFS were tested. We found that breadth but not expression level (measured as peak and mean expression) is consistently correlated with both AS and GFS whereas peak and mean expression are only marginally related with AS and GFS in some species (Figure 3). Accordingly, analyses presented below focus on expression breadth only.

Figure 3: Contributions of the three measures of expression in predicting AS and GFS. The left panel shows R-squared of stepwise between AS and breadth, peak and mean of gene expression. Breadth shows largest proportion of R-squared. The right panel presents R-squared of stepwise between GFS and breadth, peak and mean of gene expression. Breadth shows largest proportion of R-squared.

To test whether the relationship between AS and GFS could be explained by the link of both variables with breadth, we performed forward stepwise analysis. We found gene expression breadth to be the best predictor for both AS and GFS. Notably, AS was not included as a relevant variable in the stepwise models predicting GFS in any of the nine species examined whereas GFS was included as a relevant variable with a marginal contribution in the stepwise tests as a predictor of AS for only three of the nine species

analysed (Figure 4). These results suggest that the significant and consistent covariance of breadth of expression with both AS and GFS largely accounts for previously reported inverse correlations between both AS and GFS. Similar results were obtained when using three other expression datasets covering some species: BodyMap-Xs (EST based expression data (Ogasawara et al. 2006)), Microarray data (Su et al. 2004) and SAGE data (NCBI) (supplementary Figure S2-S4).

Figure 4: Gene expression breadth is the better predictor in the relation between AS and GFS. The right panel shows R-squared of stepwise between AS and breadth and GFS. Breadth shows larger proportion of R-squared than that in GFS. The left panel presents R-squared of stepwise between GFS and breadth and AS. Breadth shows larger proportion of R-squared than that in AS.

3.4 Discussion

Here we have showed that the previously reported inverse correlation between gene family size and alternative splicing (Kopelman et al. 2005; Su et al. 2006; Jin et al. 2008) is not universal but constrained to some species regardless of whether the analyses is constrained to multi-gene families or not. In addition, no support was found for the previously reported lower rate of AS for singleton genes compared to multi-gene families (Jin et al. 2008; Roux and Robinson-Rechavi 2011). These observations remain unchanged even after correcting for variations in transcript coverage among genes and between species known to have a strong impact in AS detection (Brett et al. 2002; Kim et al. 2007a; Nilsen and Graveley 2010).

In contrast, the relationship between AS and GSF with breadth of expression is consistently found in all species analysed with expression intensity marginally contributing to predicting AS and GFS in most species. We further found that any covariance between AS and GFS is largely explained as a by-product of the relationship of both variables with expression breadth.

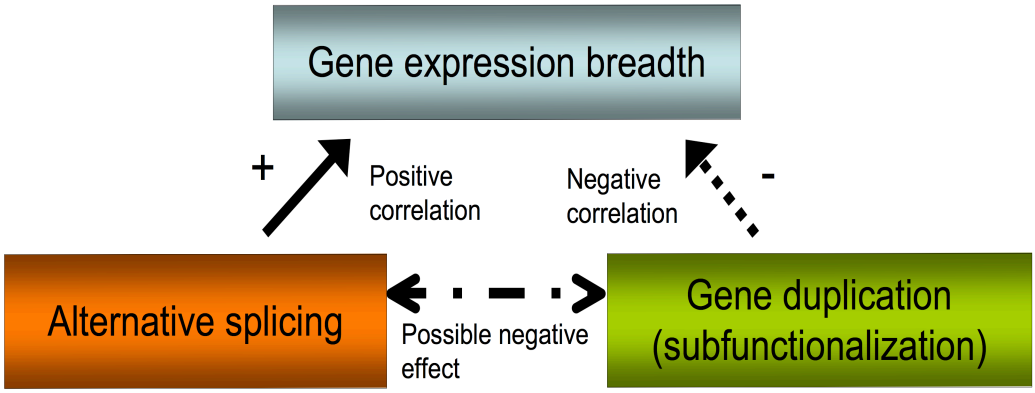


Figure 5: The relationship among gene expression breadth, alternative splicing and gene duplication. Consistent positive correlation were found between AS and breadth, in contrast, negative correlation were shown between breadth and gene duplication. Which possibly explain the negative effect between AS and gene duplication which leads to the increase of GFS.

Over time, genes acquire novel alternative splicing isoforms (Kim et al. 2007a; Wegmann et al. 2008; Roux and Robinson-Rechavi 2011; Warnefors and Eyre-Walker 2011) allowing them to specialize their function when expanding their expression to new tissues resulting in a positive correlation between AS and breadth of expression. Gene duplication events, in contrast, are often followed by the subfunctionalisation of both copies with each being expressed in fewer tissues than the ancestral single copy (Huminiecki and Wolfe 2004; Freilich et al. 2006; Farre and Alba 2010) and resulting in a negative correlation between gene family size and breadth of expression (Figure 5). We can expect a negative correlation between alternative splicing and gene family size only where a stable optimum 'transcript diversity' level exists. An expanding number of proteins throughout evolution for any given family can result from either an expansion in gene number, an expansion in alternative splicing isoforms or by an expansion of both. To test whether the relationship between AS and GFS differs for gene families with or without expanding number of transcripts throughout evolution, we calculated overall transcript diversity in over 3000 gene families (see methods) and divided them in three groups: expanded, stable and decreased transcript diversity. We found that those families with stable or decreased transcript diversity over time, exhibit a negative correlation between AS and GFS whereas for those gene families with an increased transcript

diversity the two variables are positively related (Figure 6). Interestingly, we also found that gene families with expanding, stable and decreasing transcript diversity are not distributed over all functional categories equally. Gene functions associated with cell-to-cell communication, development and behavior are associated with increased transcript diversity (Figure 7).

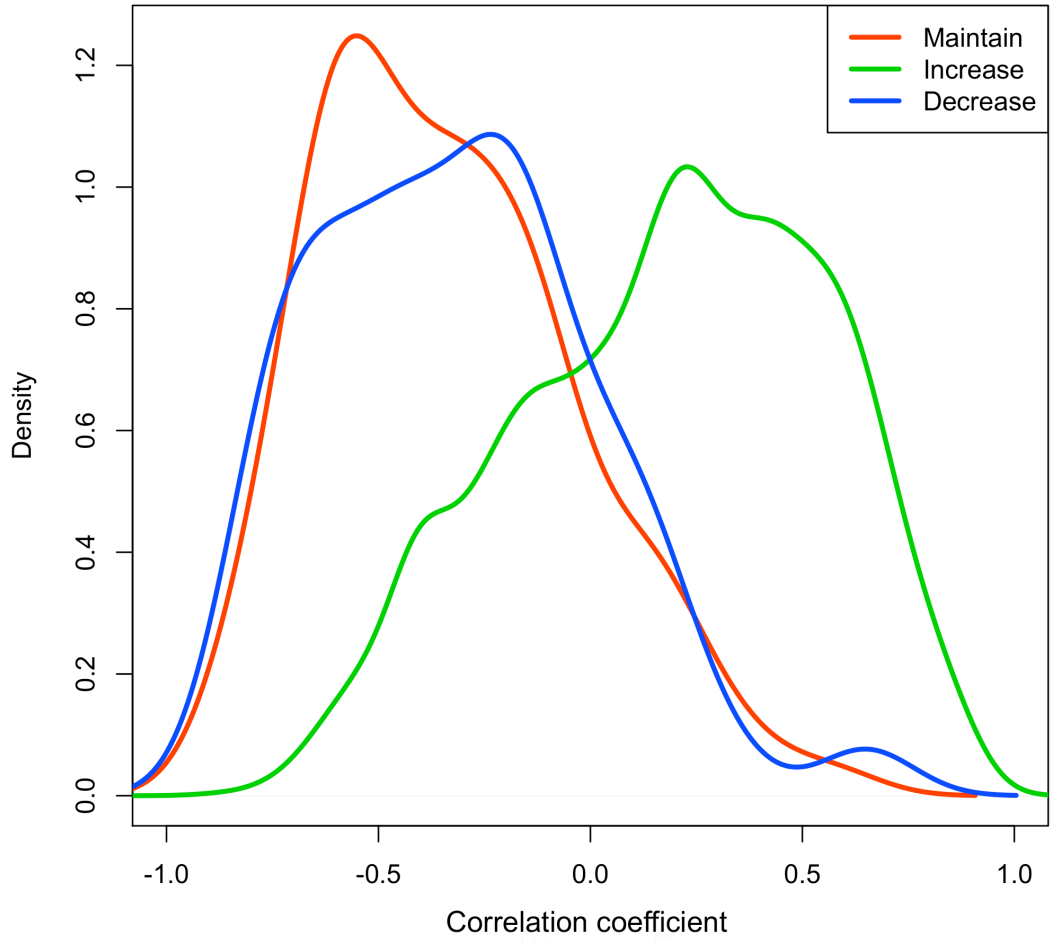


Figure 6: The distribution of correlation coefficient of AS vs. GFS in three groups. The gene families were divided into different groups according to three fates of transcript diversity of gene family including maintain, increase and decrease.

We conclude that alternative splicing and gene duplication far from being mutually exclusive mechanisms, are joint contributors of transcript diversity within gene families in close association with gene expression breadth.

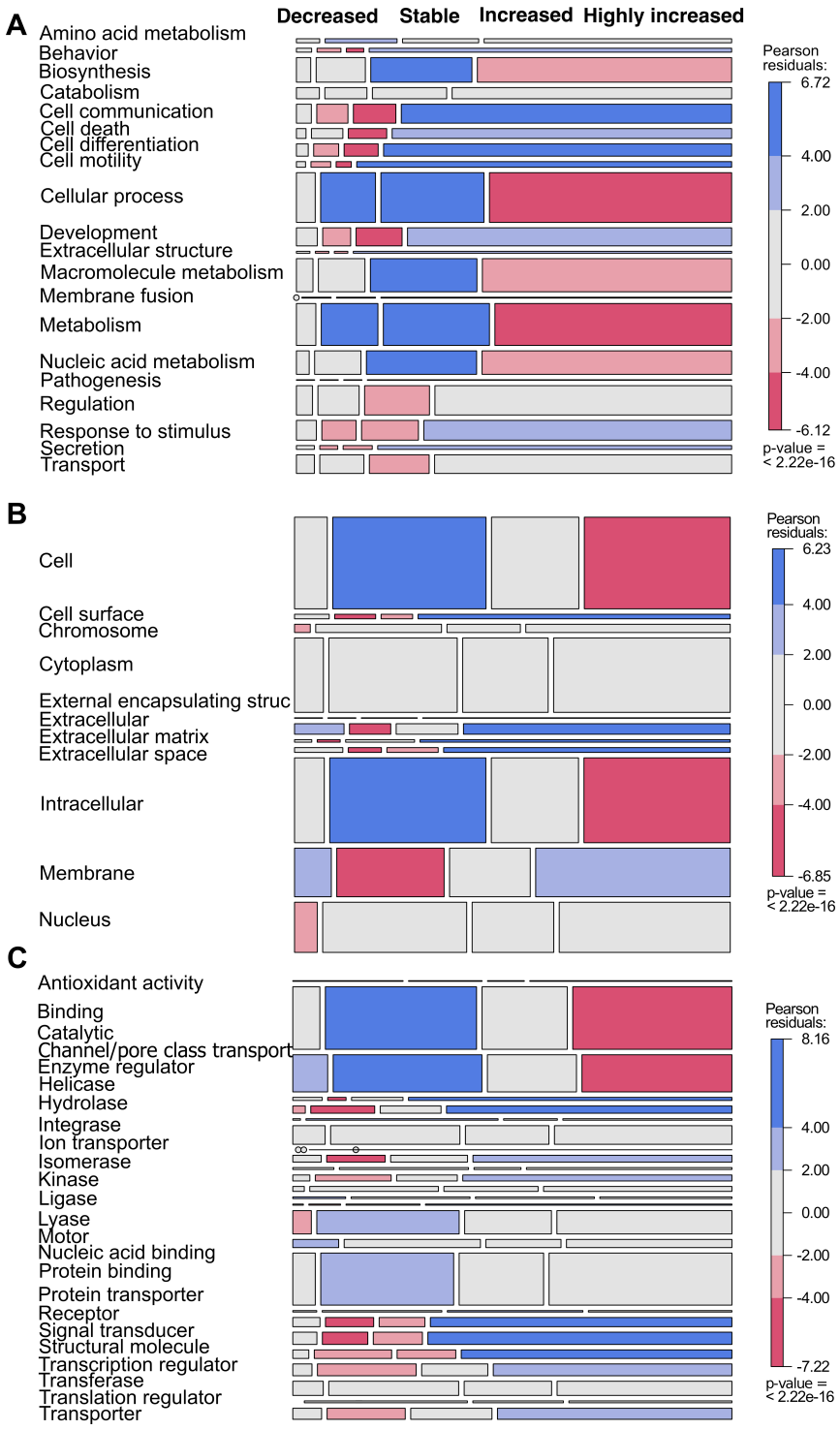


Figure 5. Functional heterogeneity in proteome expansion. Relation between GO categories and changes in transcript diversity comparing 3879 gene families (presented in at least three invertebrate and three vertebrate species). Each row represents the gene families associated with each functional category; the height of each row represents the proportion of families associated with each functional category. Colours denote whether, for any given functional category, the number of gene families in a particular group or block is above expectations (blue), under expectations (red) or at expected levels (grey). Mosaic plots were adopted according to the webpage (http://www.statmethods.net/advgraphs/mosaic.html). The Chi-

squared test was used for test the proportion among GO categories from four groups.

3.5 Supplementary Materials

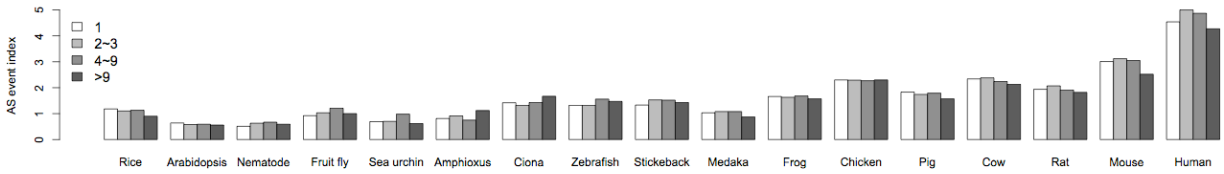


Figure S1. The comparison between the gene family size and the AS event index. Histogram bars indicate the randomized AS index, classified as singletons (1 member) and small (2~3 members), medium (4~9 members) and large (>9 members) gene families.

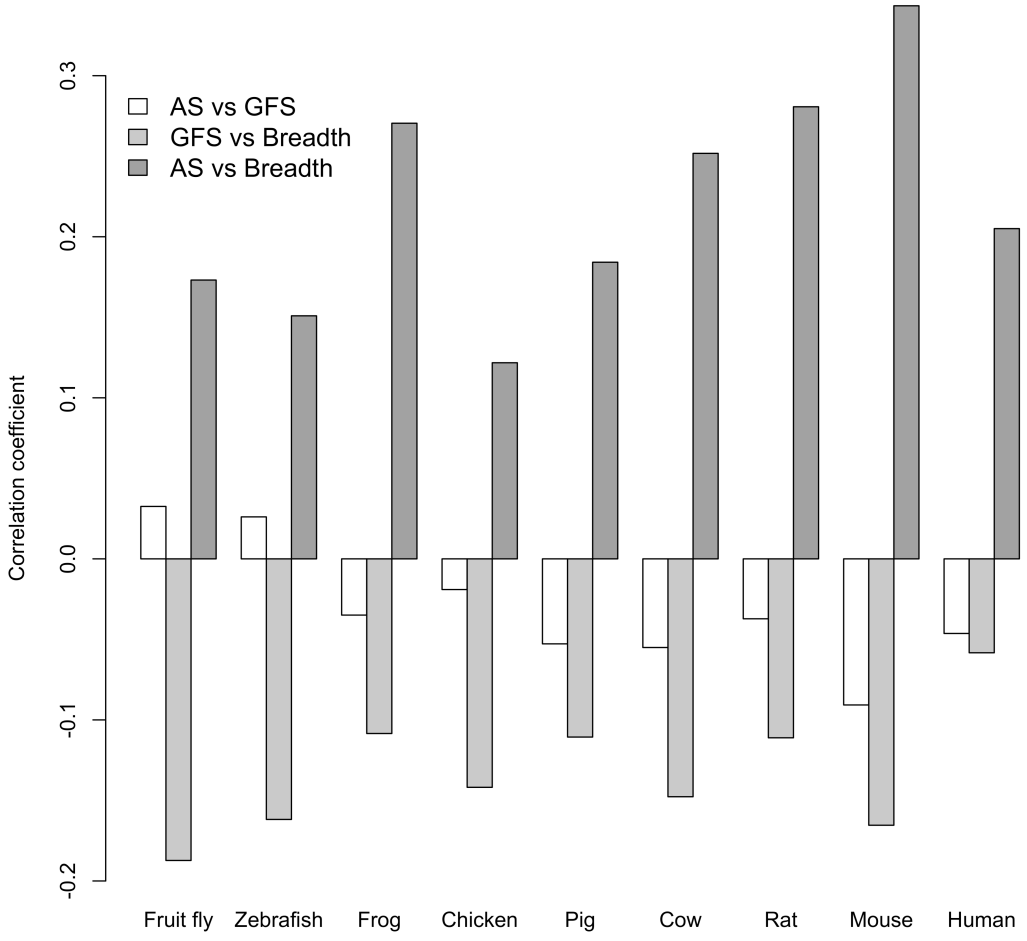


Figure S2: Comparison of the relationship between AS, GFS and breadth. The histogram bars indicate the correlation coefficient of AS versus GFS, GFS versus Breadth, AS versus breadth in nine species from BodyMap-Xs, respectively.

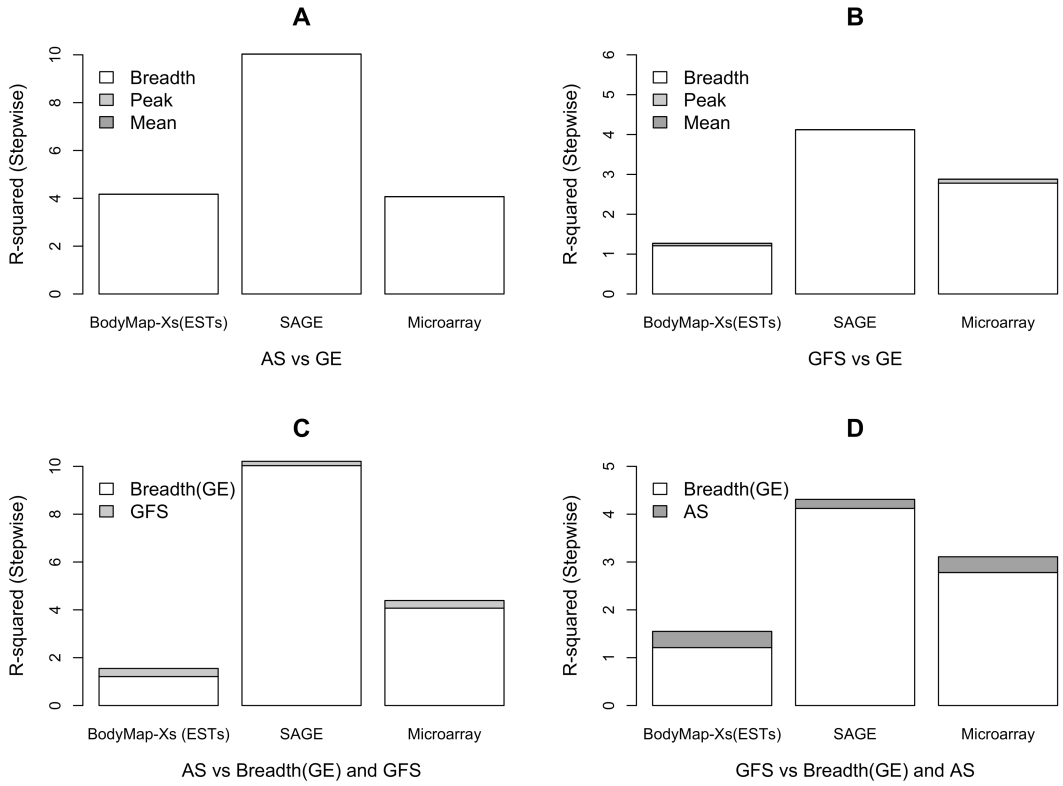


Figure S3: Stepwise tests of AS, GFS and GE using BodyMap-Xs, SAGE and Microarray data in human. (A) R-squared of stepwise between AS and breadth, peak and mean of gene expression (GE). Breadth shows largest proportion of R-squared. (B) R-squared of stepwise between GFS and breadth, peak and mean of gene expression. Breadth shows largest proportion of R-squared. (C) R-squared of stepwise between AS and breadth and GFS. Breadth shows larger proportion of R-squared than that in GFS. (D) R-squared of stepwise between GFS and breadth and AS. Breadth shows larger proportion of R-squared than that in AS.

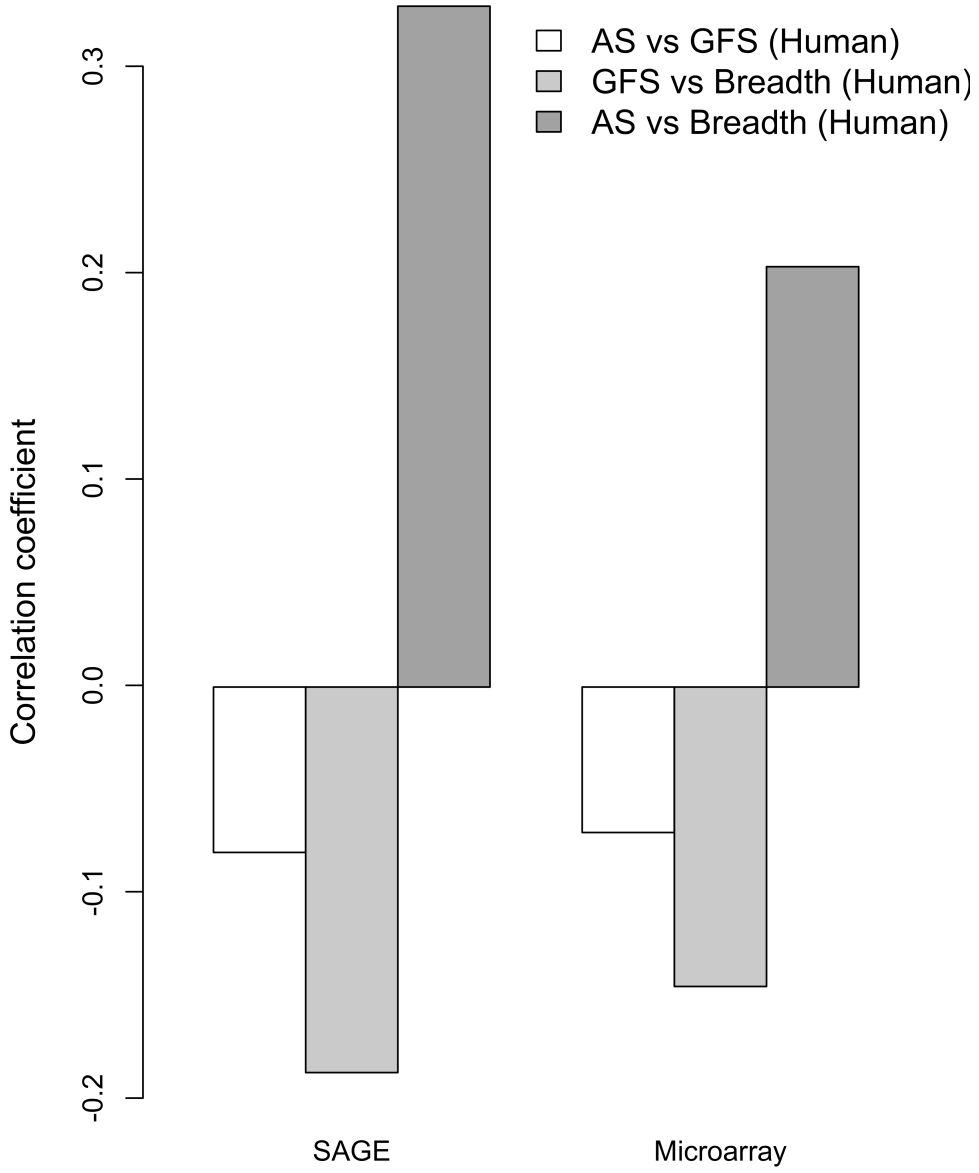


Figure S4. Comparison of the relationship between AS, GFS and breadth. The histogram bars indicate the correlation coefficient of AS versus GFS, GFS versus breadth, AS versus breadth in human using BodyMap-Xs, SAGE and Microarray, respectively.

Table S1: Paper summary for the relationship between AS and GFS

Species	Data	Alternative Splicing	Gene family size	Bias control	Correlation	Ref.
Human	Ensembl	ASD's AltSplice database	BLSATP	Exons, EST coverage, gene family size, isoform count	Negative correlation, decreased AS percentage	Kopelman et al. 2005
	NCBI, UCSC	GeneSplicer program	EnsMart	Remove garbage EST, EST coverage,	Negative correlation, increased no AS percentage	Su et al. 2006
	H-InvDB 5.0	H-InvDB 5.0	BLAST		Duplicates with higher AS percentage, decreased AS percentage in duplicates	Jin et al. 2008
Mouse	Ensembl	ASD's AltSplice database	BLSATP	exons, EST coverage, gene family size, isoform count	Negative correlation, decreased AS percentage	Kopelman et al 2005
	NCBI, UCSC	GeneSplicer program	EnsMart	Remove garbage EST, EST coverage,	Negative correlation, increased no AS percentage	Su et al. 2006
	Riken's FANTOM3	Riken's FANTOM3	BLAST		Duplicates with higher AS percentage, decreased AS percentage in duplicates	Jin et al. 2008
C.elegans	WormPep	WormPep	BLASTCLUST		Significant difference, decreased AS percentage	Hughes & Friedman 2008
Rice	TIGR 4.0	PASA program	Pfam HMM & BLASTP-based domains	Remove genes that lack transcript evidence	Multi-gene family have more AS percentage than singletons with Significant difference	Lin et al. 2008
Arabidopsis	TAIR7	TAIR7	TAIR7		Multi-gene family have more AS percentage than singletons with significant difference	Lin et al. 2008

Table S2: The source of genomes, predicted genes and ESTs.

Tax ID	Species	Genome & Predict gene	Version
4530	*Oryza sativa*	Ensembl	MSU6
3702	*Arabidopsis thaliana*	Ensembl	TAIR9
6239	*Caenorhabditis elegans*	Ensembl	WS190.54
7227	*Drosophila melanogaster*	Ensembl	BDGP5.4.54
7668	*Strongylocentrotus purpuratus*	NCBI	Spur_2.1
7739	*Branchiostoma floridae*	JGI	JGI2.0
7719	*Ciona intestinalis*	Ensembl	JGI2.55
7955	*Danio rerio*	Ensembl	Zv8.55
69293	*Gasterosteus aculeatus*	Ensembl	BROADS1
8090	*Oryzias latipes*	Ensembl	HdrR
8364	*Xenopus tropicalis*	Ensembl	JGI4.1.54
9031	*Gallus gallus*	Ensembl	WASHUC2.54
9823	*Sus scrofa*	Ensembl	Sscrofa9
9913	*Bos taurus*	Ensembl	Btau_4.0
10116	*Rattus norvegicus*	Ensembl	RGSC3.4.55
10090	*Mus musculus*	Ensembl	NCBIM37
9606	*Homo sapiens*	Ensembl	NCBI36.54

Table S3 Comparison of AS occurrence between singletons and duplicates

Organism	Singleton gene	AS percentage	Average AS	Duplicate gene	AS percentage	Average AS	*P* value (Mann-Whitney test)
Oryza sativa	2608	50.12%	1.18	11008	44.61%	1.02	3.44E-05
Arabidopsis thaliana	2289	33.16%	0.64	10151	29.64%	0.57	0.0370
Caenorhabditis elegans	1951	24.30%	0.51	2568	26.48%	0.63	0.0012
Drosophila melanogaster	2400	38.92%	0.92	3012	40.77%	1.08	0.0052
Strongylocentrotus purpuratus	546	21.98%	0.69	493	23.53%	0.77	0.1402
Branchiostoma floridae	803	32.00%	0.81	577	32.41%	0.94	0.8614
Ciona intestinalis	2192	51.82%	1.42	2925	52.17%	1.41	0.8312
Danio rerio	2260	54.96%	1.32	6385	54.46%	1.45	0.0464
Gasterosteus aculeatus	452	55.09%	1.33	1290	58.91%	1.50	0.0904
Oryzias latipes	1000	51.20%	1.03	1817	48.65%	1.03	0.6083
Xenopus tropicalis	1705	63.23%	1.66	3560	57.95%	1.63	0.0663
Gallus gallus	2061	73.22%	2.30	3438	70.45%	2.29	0.2468
Sus scrofa	1936	66.12%	1.83	3345	59.79%	1.72	7.20E-05
Bos taurus	2626	74.07%	2.34	5794	70.61%	2.27	0.0084
Rattus norvegicus	2252	65.59%	1.94	5413	64.66%	1.95	0.8630
Mus musculus	3915	79.05%	3.01	9502	77.38%	2.91	0.0140
Homo sapiens	3992	85.10%	4.54	9298	87.65%	4.73	0.0046

Table S4. Correlation test between gene family size and AS occurrence using randomized AS.

Organism	All Genes	*P*	*R*	Duplicates	*P*	*R*
Oryza sativa	13616	3.73E-09	-0.0505	11008	6.04E-13	-0.0685
Arabidopsis thaliana	12440	0.0629	-0.0167	10151	0.0872	-0.0170
Caenorhabditis elegans	4519	0.0034	0.0435	2568	0.2422	-0.0231
Drosophila melanogaster	5412	0.1935	0.0177	3012	0.0011	-0.0597
Strongylocentrotus purpuratus	1039	0.1211	0.0481	493	0.6704	0.0192
Branchiostoma floridae	1380	0.1747	0.0366	577	0.7317	0.0143
Ciona intestinalis	5117	0.5466	0.0084	2925	0.0079	0.0491
Danio rerio	8645	0.0286	0.0235	6385	0.8371	-0.0026
Gasterosteus aculeatus	1742	0.1773	0.0323	1290	0.9708	0.0010
Oryzias latipes	2817	0.1125	-0.0299	1817	0.0011	-0.0765
Xenopus tropicalis	5265	0.1127	-0.0219	3560	0.0132	-0.0415
Gallus gallus	5499	0.9370	-0.0011	3438	0.6009	-0.0089
Sus scrofa	5281	0.0135	-0.0340	3345	0.1464	-0.0251
Bos taurus	8420	0.0108	-0.0278	5794	6.90E-05	-0.0523
Rattus norvegicus	7665	0.5257	-0.0072	5413	2.42E-05	-0.0574
Mus musculus	13417	1.94E-05	-0.0369	9502	3.82E-32	-0.1206
Homo sapiens	13290	0.0015	-0.0276	9298	8.38E-28	-0.1130

4 Transcript diversification by gene duplication and alternative splicing accounts for complexity increases over eukaryotic evolution

4.1 Introduction

Despite two rounds of genome duplication at the base of the vertebrate lineage (Ohno 1970; Dehal and Boore 2005), our genome contains almost as many genes as a worm (Lander et al. 2001). Alternative splicing (AS), is a post-transcriptional process in eukaryotic organisms by which multiple distinct transcripts are produced from a single gene, and as such it has the potential to boost the total number of distinct proteins encoded in a genome (Nilsen and Graveley 2010). Recent deep sequencing analyses in the human transcriptome have shown that over 90% of multi-exon genes undergo alternative splicing (Pan et al. 2008; Wang et al. 2008) and AS has been proposed as a potential determinant of organism complexity (Xing and Lee 2006). Efforts to assess alternative splicing variation among species have resulted in conflicting results partly because of the large differences in transcript coverage between genes and organisms (Brett et al. 2002; Heebal Kim 2004; Kim et al. 2007a; Takeda et al. 2008; Mollet et al. 2010). Thus, the contribution of AS to transcript diversity and complexity throughout evolution remains unknown (Nilsen and Graveley 2010).

Here we assess the prevalence of AS in 18 eukaryotic genomes that have diverged from the lineage leading to humans over the last 1.4 billion years. We then estimated overall transcript diversity to examine how it relates to organism complexity and other previously described genomic correlates of complexity.

4.2 Materials and methods

4.2.1 Data sources

Genome sequences and annotations were obtained from sources in Table S1. Full mRNA and EST sequences were downloaded from UniGene (Sayers et al. 2009). Cancer derived EST libraries for human and mouse were removed from all analyses presented.

4.2.2 Identification of alternative splice events

To estimate AS events in different organisms, a novel procedure was applied as follows:

(1) *Mapping predicted genes and ESTs to Genome and grouping ESTs for each gene.* Overlapping and nested genes were identified and removed from further analyses. GMAP (Wu and Watanabe 2005) was used to align full transcripts and high quality ESTs to their corresponding predicted genes. Genes with no matching transcript were removed.

(2) *Template building*. To obtain an exon template as complete as possible (as well as overcoming the fact that some invertebrates do not have full transcripts sequenced) full transcripts and ESTs were overlaid onto the genomic sequence (Figure S1). This was done as follows: First the longest partial or full transcript available forms the base of the template. All other mRNAs and ESTs are then aligned to the genomic sequence and boundaries with the previously included transcripts are revised to extend exons or include new ones. If a transcript only encompasses a single exon then it will be discarded. This allows identifying and discarding any single exon nesting genes which have not been previously annotated.

(3) *Detecting AS events.* We developed an algorithm to compares the exon boundaries of any transcript to its corresponding template. Discrepancies of less than 15 bp in length were discarded. We identified eight types of AS events (Figure S2).

(4) *Obtaining comparable AS data across genes and species.* In order to avoid coverage biases, one hundred samples of 10 randomly selected transcripts were obtained per gene in every species. AS levels and isoform number were then calculated as described above in each sample and results were averaged per gene.

(5) *Identification of AS isoforms.* To identify AS isoforms, transcripts were first sorted according to the number of AS events they contain. Then transcripts containing identical or similar AS events were classed as redundant and excluded from the analysis. The number of remaining transcripts was taken as estimate of AS isoforms produced per gene.

(6) *Estimating total number of isoforms produced per gene.* AS isoforms were calculated for all genes with over 100 transcripts available. These numbers were then correlated with those obtained from 10 transcript samples. The resulting regression equation for each species was then used to extrapolate isoform number from the 10 transcript samples to an estimated number of total isoform number produced per gene

(Figure S3 and Table S3). For species with less than 100 genes with over 100 transcripts, the equation of a close relative was used instead: ciona's for sea urchin and amphioxus, medaka's for stickleback genes. As chicken AS levels are closer to those of mammals, lizard's AS was predicted with the zebrafish equation. Transcript diversity was estimated by multiplying average isoform number per gene with total gene number in each species after removing isoforms with internal stop codons.

4.2.3 Identification of paralogs and orthologs

Orthology and paralogy information was obtained from BioMart (Haider et al. 2009). For sea urchin and amphioxus, BLASTP (Altschul et al. 1997) and InParanoid (Berglund et al. 2008) were used to assemble gene families and reconstruct orthology relations. We defined 3879 gene families with one or more genes that were present in at least three invertebrate and three vertebrate species. For randomization protocol, we used the number of gene family size, AS isoform and transcript diversity from these gene families, and run linear regression between the random value (gene family size, AS and transcript diversity, respectively) against divergence time (1000 times) in order to show their trends (slope of the regression) through the time.

4.2.4 Function and structure prediction of AS isoform.

To calculate the proportion of AS transcripts with stop codons, BLASTX (Altschul et al. 1997) was run to search transcripts ORF according to protein sequences, we then deduced amino acid sequences for each AS isoform. From the BLASTX alignment files, we further extracted amino acid sequences of AS area and stop codons were identified. As the levels of stop codons vary greatly even between closely related species due to differences in sequencing quality, stop codon presence in AS areas of transcripts were corrected by the number of stop codons in constitutively translated areas.

To evaluate and characterize the functions and structure of AS, we used InterProscan which contains 14 applications for the prediction of protein domains (Zdobnov and Apweiler 2001), including Pfam for the prediction of protein domains (Bateman et al. 2004), SignalP 3.0 for signal peptide predictions (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001) for the predictions of trans-membrane domains. PSORT II (Nakai and Horton 1999) was conducted for the sub-cellular localization signal predictions. Secondary protein structures were predicted by CLC Main Workbench 5.7,

which is based on extracted protein sequences from the protein databank (http://www.rcsb.org/pdb/).

4.2.5 Intergenic space, average intron length, TE content and recombination rates

For the relationship between intergenic space and AS, we adapted the method to calculate the intergenic space from a previous study (Nelson et al. 2004). Genome size and average intron length per gene were obtained from the gene information from Ensembl. Transposable elements data were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/).

4.3 Results and discussion

4.3.1 AS prevalence has increased throughout evolution

In order to assess whether alternative splicing levels have changed over time, we identified AS events for each gene in 18 eukaryotic genomes (species listed in Table S1) from all partial and full transcripts available (cancer libraries and other diseased tissue libraries available were removed from further analyses). To this end, all available transcripts were aligned to each gene. Using these alignments a full exon intron gene template was constructed resulting in the identification of previously un-annotated exons in all species analysed. Orphan exons not supported by any transcript aligned to any other exon in the gene were removed, as they are likely to represent unannotated exons of overlapping genes or nested single exon genes. All transcripts containing premature stop codons were also removed from further analyses. To minimise the strong dependence of AS detection on transcript coverage per gene (Brett et al. 2002; Kim et al. 2007a; Nilsen and Graveley 2010), we used a randomisation protocol (adapted from (Kim et al. 2007a); see methods) identifying alternative splicing events in 1000 samples of 10 transcripts. Genes with less than 10 transcripts were removed from further analyses. Our results show that if species are arranged according to divergence time from the lineage leading to humans (data from ref. (Hedges et al. 2006)) AS levels and the percentage of genes with at least one AS event detected (average from ten transcript samples) have increased over the last 1.4 billion years from virtually none in yeast to 94.8% of genes being alternatively spliced in humans (Figure 1). This increase in AS levels is consistent with

observations by Kim et al. (Kim et al. 2007a). Our results show a much higher AS prevalence in non-human species (36.5% ~68.0% for invertebrates and 68.0%~93.4% for vertebrates) than previous estimates (15% for invertebrates and 30-45% for vertebrates: (Kim et al. 2007a). Importantly, our assessment of AS levels using publicly available transcript data for human and *Drosophila melanogaster* resulted in a similar percentage of genes being identified as having alternative splicing as recent studies using high throughput sequencing technology (Pan et al. 2008; Wang et al. 2008; Graveley et al. 2011). We also reject previous findings suggesting a low AS occurrence in birds (Chacko and Ranganathan 2009) as chicken's AS levels are more similar to that of mammals than to its closer relative anolis lizard (Figure 1A and 1B and Table 1).

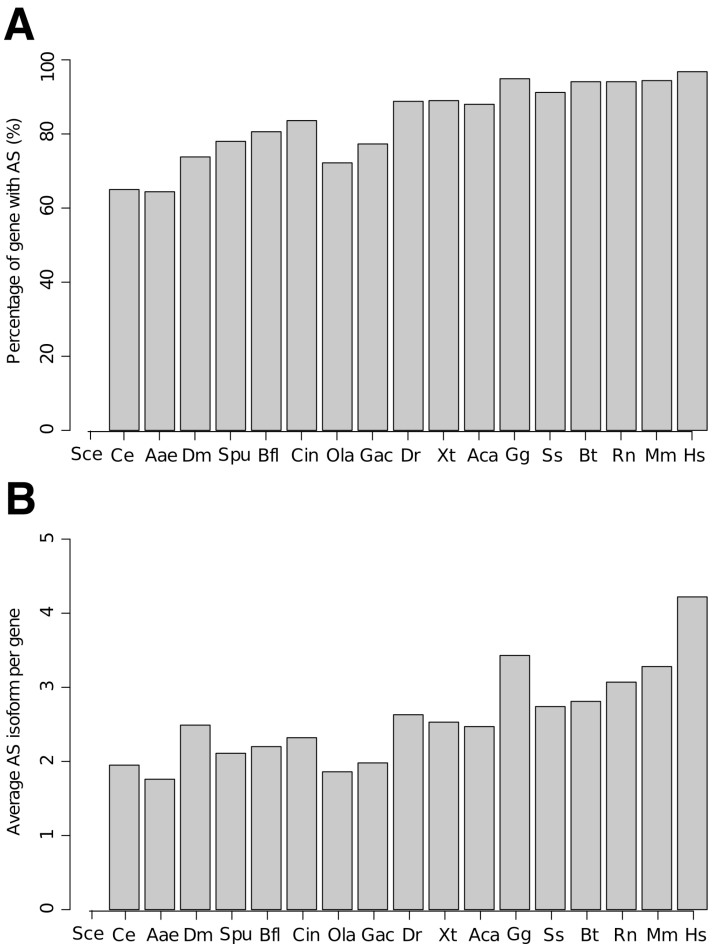


Figure 1. AS levels in 18 eukaryotic genomes and analysis of functional content. (A) AS prevalence in 18 genomes. (B) Average AS isoforms per gene. Species names are listed in Table S1.

Table 1. Summary statistics for AS analysis in 18 eukaryotic genomes.

Species	Gene #	Transcripts per gene	Genes with > 10 transcripts	AS genes (%)*	AS isoforms	AS isoforms (sampling)	Proteome size
Yeast	7000	6.8	431	-	0	1	7000
Nematode	19000	44.3	4519	42.8	0.65	1.95	35358
Mosquito	16000	36.8	4644	36.5	0.44	1.76	27463
Fruit fly	15000	78.1	5412	59.2	1.11	2.49	35892
Sea urchin	21000	28.4	986	36.7	0.82	2.11	41958
Amphioxus	21000	46.3	1361	50.2	1.25	2.20	44766
Ciona	16000	86.2	5117	68.0	1.31	2.32	36192
Medaka	21000	72.1	2817	68.0	0.91	1.86	37265
Stickeback	22000	41.4	1742	70.5	1.03	1.98	40224
Zebrafish	22000	72.3	8645	83.4	1.45	2.63	55826
Frog	20000	81.2	5265	86.0	1.51	2.53	49230
Lizard	21000	21.6	935	71.5	0.76	2.47	51556
Chicken	17000	38.3	5499	92.9	1.94	3.43	54792
Pig	22000	83.8	5281	91.6	1.65	2.74	58080
Cow	22000	67.2	8420	92.2	1.98	2.81	58759
Rat	23000	48.7	7665	90.6	1.73	3.07	67788
Mouse	23000	177.2	13417	93.4	3.31	3.28	69297
Human	23000	305.7	13290	94.8	6.62	4.22	90919

* The AS percentage is the gene with at least 1 AS event out of 10 ESTs.

4.3.2 Contribution of alternative splicing and gene duplication to the transcript diversity

Having shown that alternative splicing levels have increased along evolution we estimated the overall contribution of alternative splicing to transcript diversity per species. To this end, we first estimated the number of alternative splicing isoforms produced per gene within our sample (see methods). This AS level index was then translated into an estimate of actual AS isoform number produced per gene within a given species by extrapolating AS rates derived from genes with over 100 ESTs to the whole gene pool for that species (see methods; see Figure S3).

Overall transcript diversity was then estimated by multiplying average isoform number per gene by total gene number in each species after removing transcripts containing premature stop codons (Table 1 and Figure 2, see methods). We found that while gene number has remained relatively stable in the last one billion years (note that bony fishes possibly underwent a whole genome duplication not shared by the terrestrial lineages (Jaillon et al. 2004; Kasahara et al. 2007), the contribution of AS isoforms to the transcript pool has increased from less than 1% in yeast to around 75% in human (Figure 2). Notably, human displayed the largest transcript diversity with an estimate of 91000 distinct transcripts (Figure 2 and Table 1). Similar AS and transcript diversity increases are found when restricting the analyses to the set of 3879 orthologous gene families present in both invertebrate and vertebrate species (see methods; Figure S4). Using a randomization protocol (see methods) we find that whereas only 40% of gene families have increased their gene number by gene duplication (slope > 0), over 80% of gene families have increased their number of AS isoforms, both contribute to the expansion of transcript diversity (Figure 3).

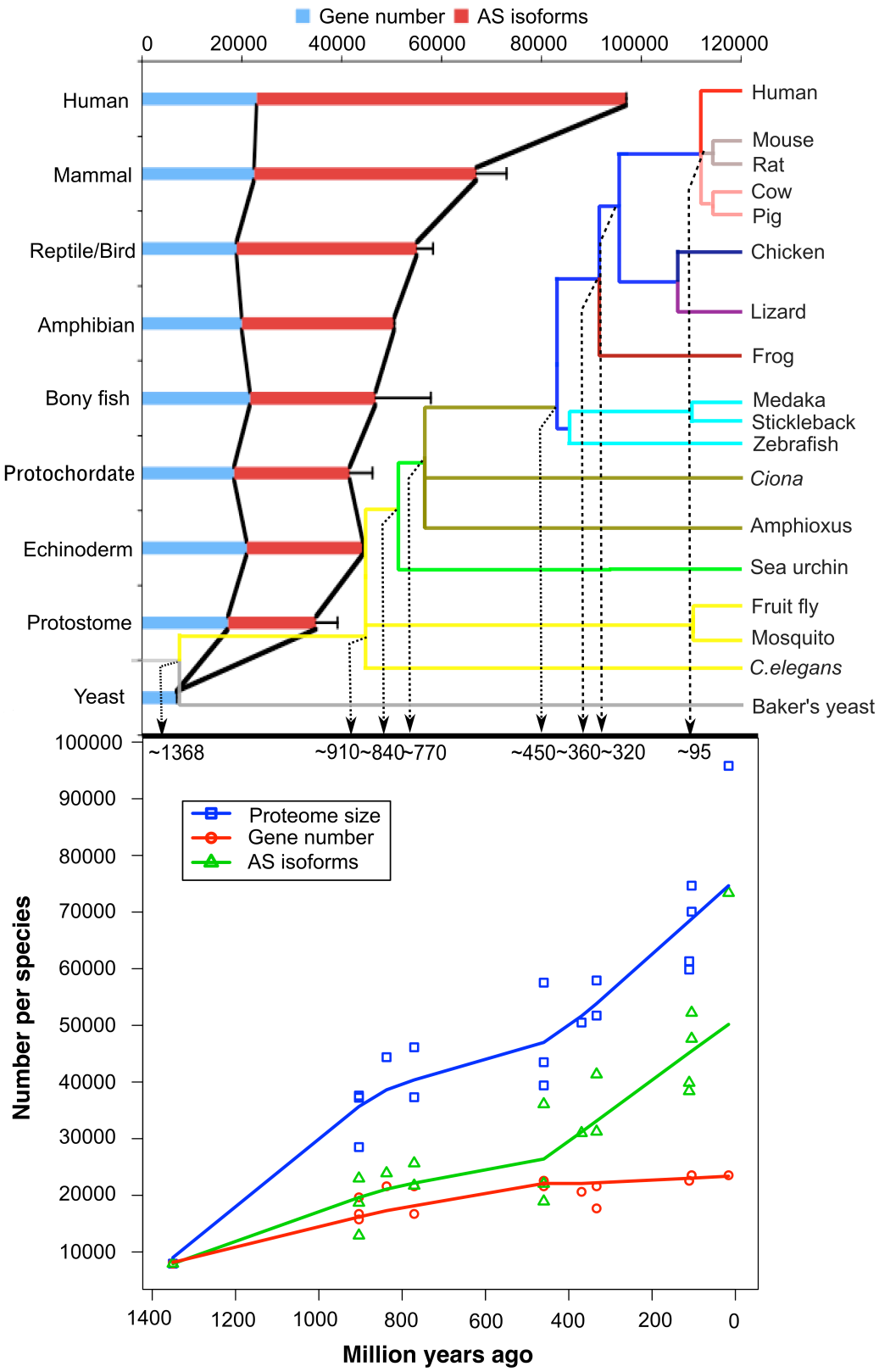


Figure 2. Evolution of gene number, AS and proteome size. Upper panel shows relative contribution of gene number and alternative splicing to proteome size for species grouped by their divergence time from the human lineage (tree adapted from (Hedges et al. 2006)). Bottom panel presents gene number, average AS isoform number and proteome size as a function of time of divergence from the lineage leading to human.

4.3.3 Transcript diversity is a strong predictor of organism complexity

Alternative splicing has been proposed to contribute to organism complexity thus explaining the relatively stable gene number since the divergence of the vertebrate lineage from protostomians (Nilsen and Graveley 2010). In order to assess the relationship between estimated transcript diversity and organism complexity we compared transcript diversity estimates against organism complexity –assayed as total cell type number per species (Valentine et al. 1994). We found that while gene number and splicing isoforms contribution to transcript diversity are significantly related with complexity (R^2 = 0.499, P = 0.0010 and R^2 = 0.696, P = 1.667e-05, respectively), total transcript diversity per species is a better predictor of cell type number (R^2 = 0.749, P = 4.372e-06; Figure 4). Similar results were obtained when analysing orthologous gene families (for gene number, R^2 = 0.582, P = 0.0006; alternative splicing R^2 = 0.611, P = 0.0001 and transcript diversity, R^2 = 0.786, P = 9.404e-07; Figure S5).

Interestingly, the larger human EST pool compared to other mammals is not reflected in a proportionally higher cell type number. This may result from the lack of resolution in cell type estimates among mammalian species or may reflect the significant higher brain to body weight ratio particular to our species of 7.6 compared to 0.5 in the other four mammalian species analysed (Roth and Dicke 2005). Future analyses will allow testing this hypothesis.

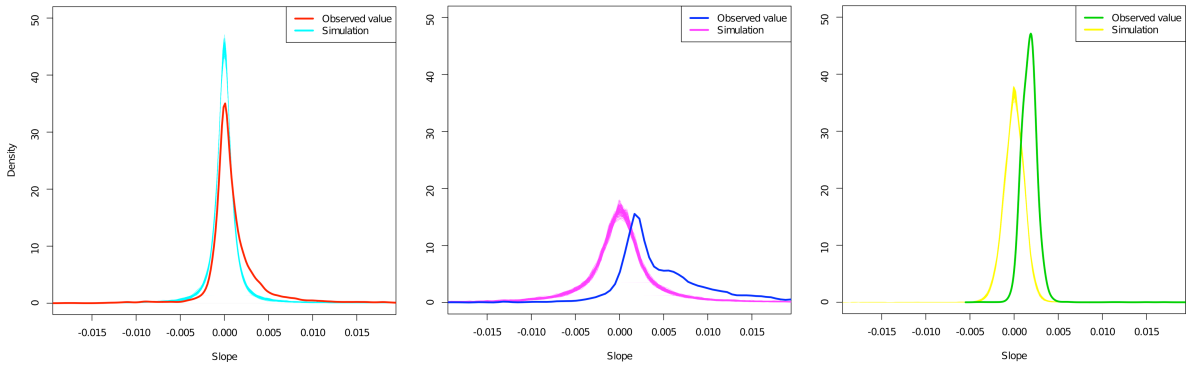


Figure 3. Simulation of the changes of gene duplication (left), transcript diversity (middle) and alternative splicing (right) in 3879 gene family through divergence time (1000 times randomization).

4.3.4 Transcript diversity is a better predictor of organism complexity than any previously reported co-varying parameters

So far we have shown that overall transcript diversity is strongly correlated with organism complexity. But how does it compare with other reported genomic predictors of organism complexity? Previous reports have shown that organism complexity is associated with proliferation of a variety of genomic features such as genome size, total gene number, intron and intergenic spacer length and transposable element (TE) content (Lynch and Conery 2003). To test the contributions of each parameter to organism complexity we used a forward stepwise regression analyses (F = 4.0) and found that transcript pool size is the only relevant predictor of complexity. Increases in functional domains have also been proposed as a contributor to complexity (Vogel and Chothia 2006; Xia et al. 2008b). We identified functional content per gene using functional component prediction software (for list see Table S2). We found that while a number of functional content parameters are related to complexity, transcript diversity explains the most variance. Using a stepwise regression analyses (F = 4.0) including all 12 functional component indexes as well as transcriptome pool size (Table S2), we found that transcript diversity is the main predictor of cell type number explaining 75% of the variance in cell type number with HMMPfam and PatternScan explaining a further R^2 = 0.0721 and 0.0486 ($P < 0.0001$). This suggests that evolution of complexity has been accompanied by increases in the number of distinct transcripts with only minor changes in the functional content of peptide sequences consistent with limitations on functional domain interference (Innan and Kondrashov 2010).

In summary, taking advantage of the increasing availability of transcript data for a variety of species and using a random sampling protocol to address biases due to differential transcript coverage we have shown that AS has increased steadily over the last 1.4 billion years. Most importantly perhaps, our estimates of transcript pool size explain over 75% of cell type number variance making it a far better predictor of complexity than any previously reported genomic predictor of complexity.

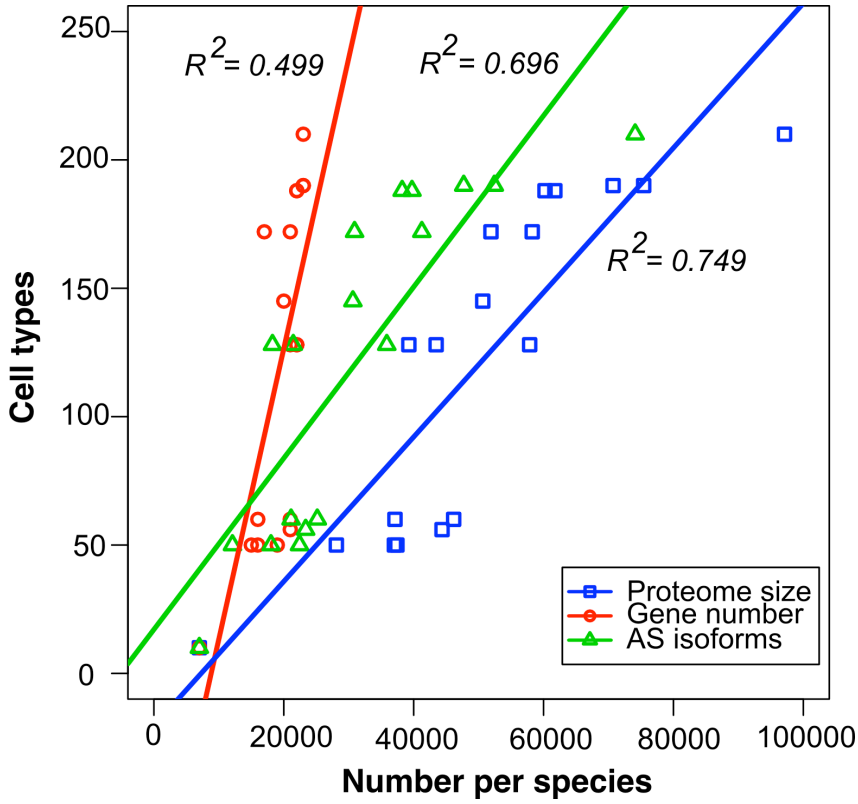


Figure 4. AS, complexity and functional gene associations. Correlation of Cell type number (data from ref. (Valentine et al. 1994)) with gene number, AS and estimated proteome size per species. Regression lines and coefficients are shown (*P* = 0.0010, 1.667 X10^{-5} and 4.372 X10^{-6} as labels appear in graph).

Table 2. Complexity, effective population size and genomic features

Million years ago	Species	Genome size	Intergenic space	Log2(Neu)*	Cell type	TE (%)	Average intron length	Average number of exon
0	Human	3.3E+09	2.1E+09	11.66	210	44	4921.9	7.4
91	Mouse	3.4E+09	1.9E+09	11.85	190	40	3721.2	6.8
91	Rat	2.5E+09	2.1E+09	-	190	40	3392.0	7.3
97	Pig	2.4E+09	1.8E+09	-	188		3677.2	8.0
97	Cow	3.2E+09	2.0E+09	-	188	27	3301.9	8.5
325	Chicken	1.1E+09	7.5E+08	-	172	9	2245.9	9.0
325	Lizard	1.7E+09	8.9E+08	-	172		3441.6	9.1
362	Frog	1.5E+09	7.6E+08	-	145	33	3016.2	7.8
455	Zebrafish	1.6E+09	8.4E+08	-	128	26	2860.4	8.6
455	Medaka	7.0E+08	5.3E+08	-	128		1428.7	9.9
455	Stickleback	4.5E+08	2.6E+08	9.95	128	2.7	880.6	9.9
774	Ciona	1.7E+08	7.1E+07	8.36	60	11	549.7	6.9
774	Amphioxus	5.8E+08	2.7E+08	-	60	30	1409.0	8.4
842	Sea urchin	8.1E+08	2.0E+08	8.76	56		1206.2	6.8
910	Mosquito	1.3E+09	7.8E+08	8.39	50	47	2553.8	3.7
910	Fruit fly	1.7E+08	1.0E+08	8.06	50	15	862.0	4.0
910	Nematode	1.0E+08	4.6E+07	8.25	50	6	233.7	5.6
1368	Yeast	1.2E+07	3.4E+06	5.45	10	3	72.3	1.0

* Neu: the product of effective population size and the mutation rate.

4.4 Supplementary Materials

Figure S1. The same with the Figure S1 in page 27.

Figure S2. The same with the Figure S2 in page 28.

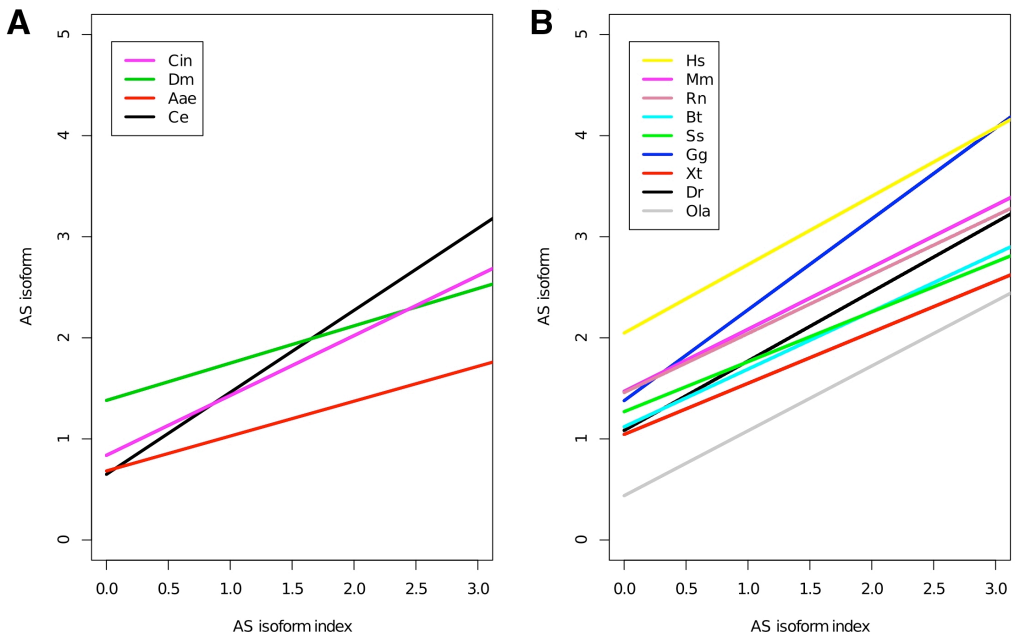


Figure S3. Linear relationship between AS index, total AS isoform number per gene for each species. Panels show the regression lines between AS index (as calculated with 10 ESTs sampling method) and AS isoform number for genes with at least 100 ESTs for invertebrate and vertebrate species.

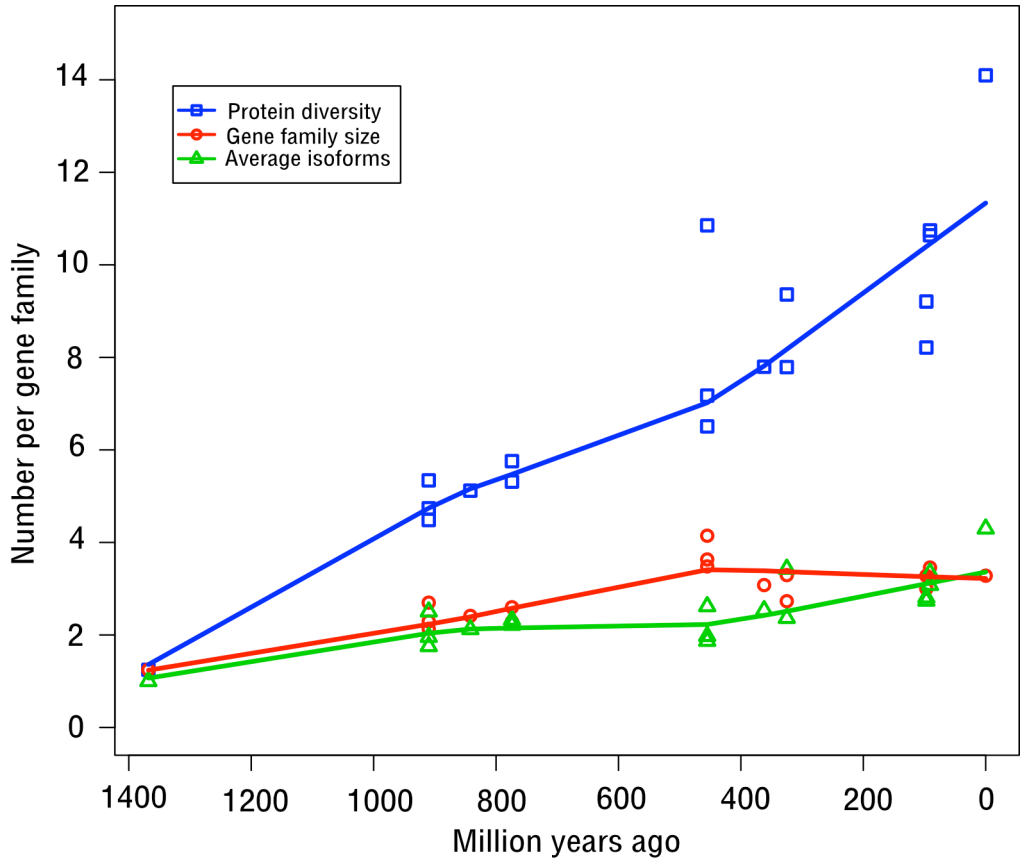


Figure S4. Evolution of alternative splicing, gene number and transcript diversity per gene family. Changes in gene number, average AS isoform number and transcript diversity in 3879 orthologous gene families present in at least three invertebrate and three vertebrate species as a function of estimated divergence time from the lineage leading to human.

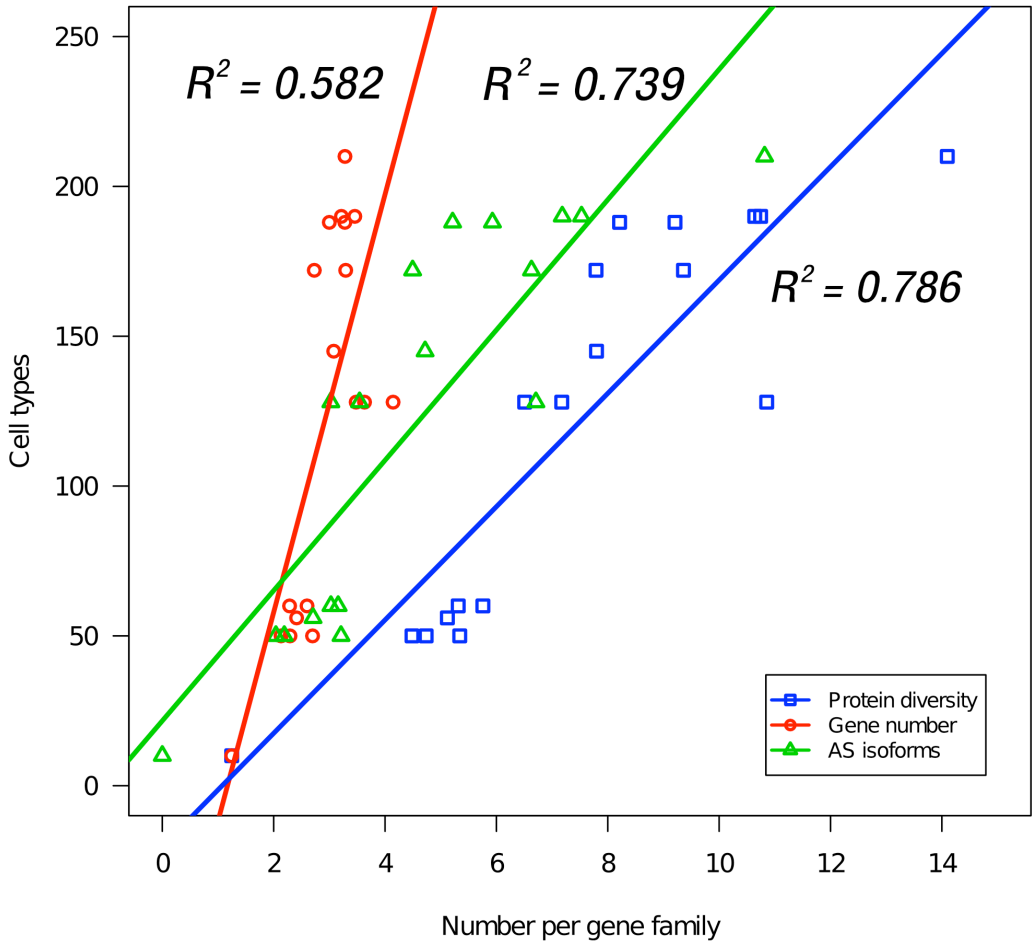


Figure S5. Relation between gene number, average AS isoforms and transcript diversity in 3879 orthologous gene families (present in at least three invertebrate and three vertebrate species). Regression lines and coefficients are shown in each panel (P = 0.0001, 1.6e-05 and 9.404e-07).

Table S1. Species list with genome and transcript data sources.

Tax ID	Species	Ab.	Genome sequence & gene annotations	Version
4932	*Saccharomyces cerevisiae*	Sce	Ensembl	SGD1.01
6239	*Caenorhabditis elegans*	Ce	Ensembl	WS190.54
7227	*Drosophila melanogaster*	Dm	Ensembl	BDGP5.4.54
7159	*Aedes aegypti*	Aae	Ensembl	AaegL1
7668	*Strongylocentrotus purpuratus*	Spu	NCBI	Spur_2.1
7739	*Branchiostoma floridae*	Bfl	JGI	JGI2.0
7719	*Ciona intestinalis*	Cin	Ensembl	JGI2.55
7955	*Danio rerio*	Dr	Ensembl	Zv8.55
69293	*Gasterosteus aculeatus*	Gac	Ensembl	BROADS1
8090	*Oryzias latipes*	Ola	Ensembl	HdrR
8364	*Xenopus tropicalis*	Xt	Ensembl	JGI4.1.54
28377	*Anolis carolinensis*	Aca	Ensembl	AnoCar1.0
9031	*Gallus gallus*	Gg	Ensembl	WASHUC2.54
9823	*Sus scrofa*	Ss	Ensembl	Sscrofa9
9913	*Bos Taurus*	Bt	Ensembl	Btau_4.0
10116	*Rattus norvegicus*	Rn	Ensembl	RGSC3.4.55
10090	*Mus musculus*	Mm	Ensembl	NCBIM37
9606	*Homo sapiens*	Hs	Ensembl	NCBI36.54

Table S2. Linear regression of average functional content per 100 amino acids versus cell type number

Application	*P*	R^2
HMMPanther	0.0058	0.4079
Gene3D	0.1459	0.1356
HMMPfam	0.0057	0.4096
Superfamily	0.0968	0.173
ProfileScan	0.4255	0.0428
HMMSmart	0.9296	0.0005
PatternScan	0.5143	0.0289
FPrintScan	0.0913	0.1783
SignalPHMM	0.8385	0.0029
TMHMM	0.595	0.0193
Secondary structure	0.2989	0.0717
Stop Codon	0.0036	0.4414

Table S3. Coefficients for regression equations of AS isoform number as a function of AS index from 10-transcript sampling.

Species	*Genes with 100~200 transcripts	AS isoform		
		a*x	b	R^2
Caenorhabditis elegans	209	0.816	0.636	0.333
Drosophila melanogaster	174	0.353	0.645	0.327
Aedes aegypti	827	0.385	1.252	0.359
Strongylocentrotus purpuratus	23	0.594	0.831	-
Branchiostoma floridae	84	0.594	0.831	-
Ciona intestinalis	580	0.594	0.831	0.509
Danio rerio	166	0.642	0.438	0.523
Gasterosteus aculeatus	88	0.642	0.438	-
Oryzias latipes	647	0.687	1.079	0.423
Xenopus tropicalis	505	0.5	1.069	0.321
Anolis carolinensis	1	0.687	1.079	-
Gallus gallus	157	0.906	1.35	0.449
Sus scrofa	528	0.499	1.246	0.329
Bos Taurus	777	0.575	1.099	0.462
Rattus norvegicus	487	0.587	1.443	0.328
Mus musculus	3326	0.62	1.441	0.406
Homo sapiens	2864	0.686	1.988	0.436

5 Cancer associated transcript quality modifications by alternative splicing

5.1 Introduction

Cancer cells are associated with profound changes at the transcriptome level with hundreds of genes being up or down regulated when compared to normal tissues (Martinez et al. 2010). Transcription profiling of cancer samples has led to an increased understanding of cancer physiology and the identification of a number of transcriptional cancer markers. Alternative splicing (AS) is a post-transcriptional process in eukaryotic organisms by which multiple distinct functional transcripts are produced from a single gene. It is now known that most human genes undergo alternative splicing (Pan et al. 2008; Wang et al. 2008). Several studies have explored cancer related changes in alternative splicing patterns (reviewed in (Kalnina et al. 2005; Venables 2006; Skotheim and Nees 2007; Wang and Cooper 2007b) resulting in the identification of an increasing number of cancer-specific AS events in a variety of cancer tissues (Xu 2003; Hui et al. 2004; Parker et al. 2004; Kim et al. 2008b; He et al. 2009). Given the high number of AS events unique to cancer transcriptomes, cancer-specific transcripts have been proposed to play a key role in cancer physiology (Skotheim and Nees 2007; He et al. 2009). Nevertheless, only a handful of cancer-specific alternative splicing events have been experimentally validated (Wang et al. 2003; Hui et al. 2004). Given that a significant proportion of alternatively spliced transcripts result from noisy splicing in normal human tissues (Green et al. 2003; Lewis et al. 2003; Zhang et al. 2009; Pickrell et al. 2010), it is possible that most cancer-specific AS result from aberrant splicing in these abnormal cells and not play any significant role in cancer onset or progression (Xu 2003; Skotheim and Nees 2007; Kim et al. 2008b). Here, by examining human and mouse EST libraries we ask whether cancer transcriptomes show any differences in transcript quality compared to normal tissues.

5.2 Materials and methods

5.2.1 Data sources

Sequence and genome annotations were obtained from Ensembl. EST sequences and library information were downloaded from UniGene (Sayers et al. 2009).

Table 1 Summary of transcripts from normal and cancer state

Species name	Disease state	Tissue type	Development stage	Library count	EST count
Human	Normal	37	7	297	1687320
	Cancer	34	5	362	920844
Mouse	Normal	29	15	164	628506
	Cancer	14	4	45	148156

5.2.2 Identification of alternative splice events

To estimate AS events in different organisms, a novel procedure was applied as follows: (i) *Mapping predicted genes and ESTs to Genome and grouping ESTs for each gene.* Overlapping and nested genes were identified and removed from further analyses. GMAP (Wu and Watanabe 2005) was used to align full transcripts and high quality ESTs to their corresponding predicted genes. Genes with no matching transcripts were removed from further analyses. (ii) *Template building*. To obtain a gene template as complete as possible, full transcripts and ESTs were overlaid onto the genomic sequence. This was done as follows: First the longest partial or full transcript available forms the base of the template. All other mRNAs and ESTs are then aligned to the genomic sequence and boundaries with the previously included transcripts are revised to extend exons or include new ones. If a transcript only encompasses a single exon then it will be discarded. This allows identifying any single exon which has not been previously annotated and discarding any non-supported exons annotated in the "predicted gene". (iii) *Detecting AS events.* We developed an algorithm for AS event detection to compare the exon boundaries of any transcript to its corresponding template. Discrepancies of less than 15 bp in length were discarded. To identify AS isoforms, transcripts were first sorted according to the number of AS events they contain. Then transcripts containing identical or similar AS events were classed as redundant. Each AS event was classified depending on whether it derives from cancer or normal libraries. Those AS events not found in either

normal or cancer libraries were deemed cancer or normal specific respectively, while AS events shared in both normal and cancer libraries were defined as normal common and cancer common respectively.

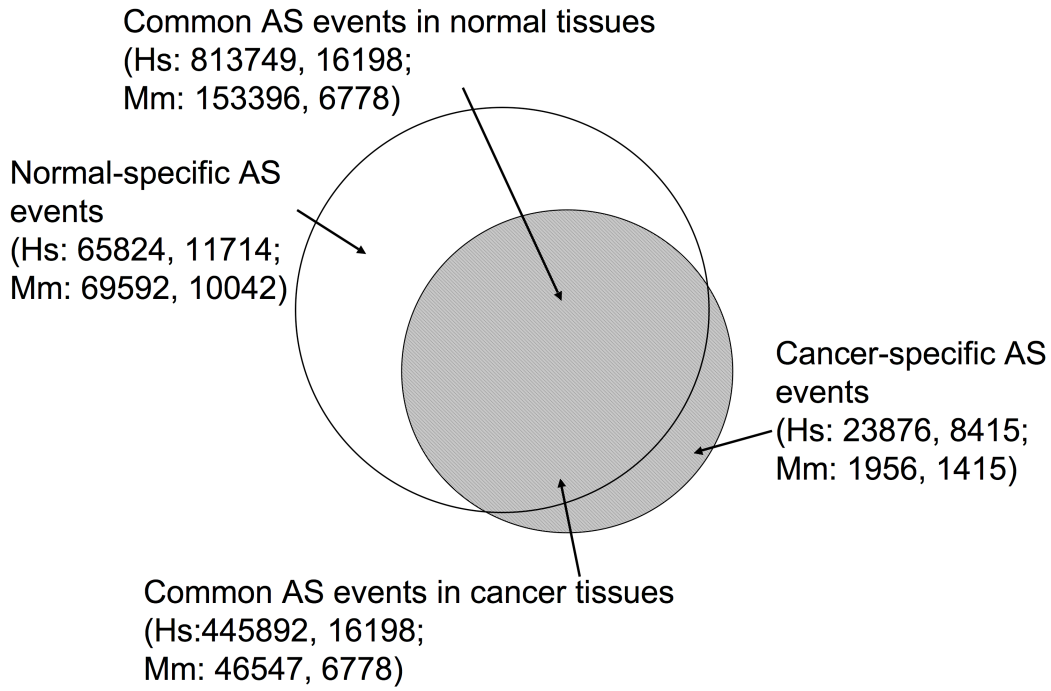


Figure 1. Schematic representation of the proportion of transcripts containing alternative splicing events common in both normal and cancer libraries, or cancer/normal specific. First number in each label represents the total number of distinct AS events detected and the second the number of genes represented for human (Hs) and mouse (Mm).

5.2.3 Identification of premature stop codons, functional and structural protein components per AS event

As transcripts supporting the same AS event may contain premature stop-codon causing mutations, stop codon presence was characterised and counted on a per transcript basis. Other features such as functional components were jointly analysed for each splicing event. To calculate the proportion of AS transcripts with stop codons, BLASTX (Altschul et al. 1997) was run to search for ORFs according to protein sequences. From the BLASTX alignment files, amino acid sequences of the AS area were extracted and stop codons were identified and counted. To functionally characterize AS events, we used

InterProScan which contains 14 applications for the prediction of protein domains (Zdobnov and Apweiler 2001), including Pfam for the prediction of protein domains (Bateman et al. 2004), SignalP 3.0 for signal peptide predictions (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001) for the predictions of transmembrane domains. PSORT II (Nakai and Horton 1999) was used to identify the likely sub-cellular localization of protein products. Secondary protein structures were predicted by CLC Main Workbench 5.7, which is based on extracted protein sequences from the protein databank (**http://www.rcsb.org/pdb/**).

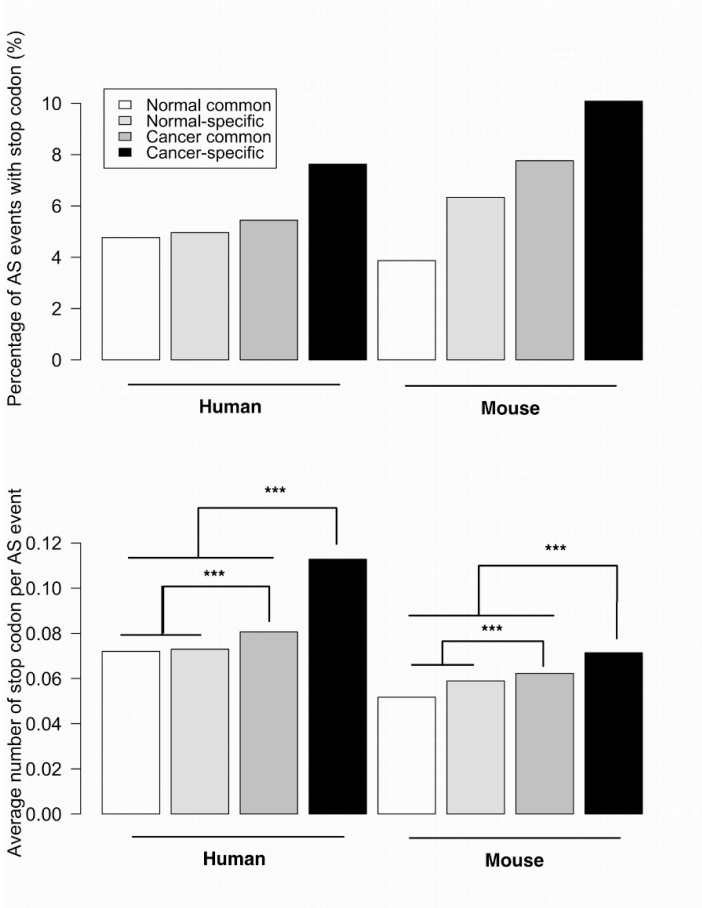


Figure 2. Premature stop codons in normal and cancer AS events. Top panel shows the percentage of premature stop codon containing AS events for normal and cancer tissues subdivided into those containing AS events unique to normal / cancer libraries or found in both. Bottom panel shows average number of premature stop codons with events divided in the same way as top panel. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with 0.01< P < 0.05 (*), 0.001< P< 0.01 (**) and P <= 0.001 (***).

5.3 Results

5.3.1 Identification of cancer-specific alternative splicing events in human and mouse

A total of 10,896,836 ESTs for human and mouse were downloaded from UniGene (Sayers et al. 2009). Of these 3,384,826 ESTs had a clear disease state annotation and were split into 297 libraries representing normal 37 tissues and 362 cancer libraries for 34 tissues for human, 164 normal libraries corresponding to 29 normal tissues and 45 cancer libraries from 14 tissues for mouse (see Table 1). To identify alternative splicing events, a complete exon template was constructed for each gene by mapping all partial and full transcripts available (using Gmap software (Wu and Watanabe 2005)). Known nested genes as well as orphan exons, not present in any transcript extending beyond them, were removed from further analysis. Individual ESTs were then aligned to the resulting gene template to identify AS events. We identified a total of 1,349,341 and 271,491 AS transcripts containing AS events for human and mouse respectively. Of these, a total of 1,259,641 (93.3%) and 199,943 (73.6%) for human and mouse respectively were found in both normal and cancer libraries while 23,876 (1.8%) and 1,956 (0.7%) were found only in cancer libraries. The remainder 65,824 (4.9%) and 69,592 (25.6%) transcripts were found to contain AS events exclusive to normal tissue derived libraries (Figure 1). The higher percentage of normal-specific AS events in mouse is explained by the limited cancer transcripts available for this species (Table 1).

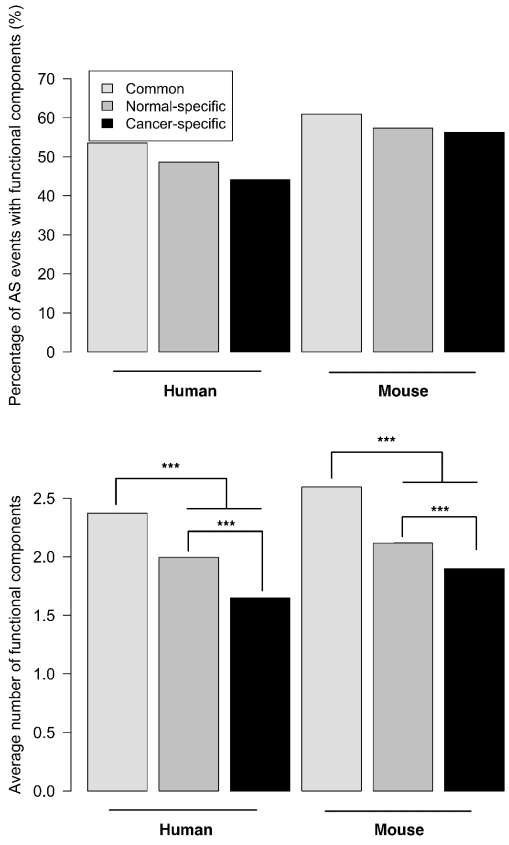


Figure 3. Identifiable functional components in AS events in cancer and normal tissues. Top panel shows the percentage of AS events with at least one identifiable functional component (see methods). Bottom panel shows average number of identifiable functional components per AS area. In both panels transcripts were divided as in Figure 2. Stars represent significant differences among groups from Wilcoxon tests with 0.01< P < =0.05 (*), 0.001< P<= 0.01 (**) and P <= 0.001 (***).

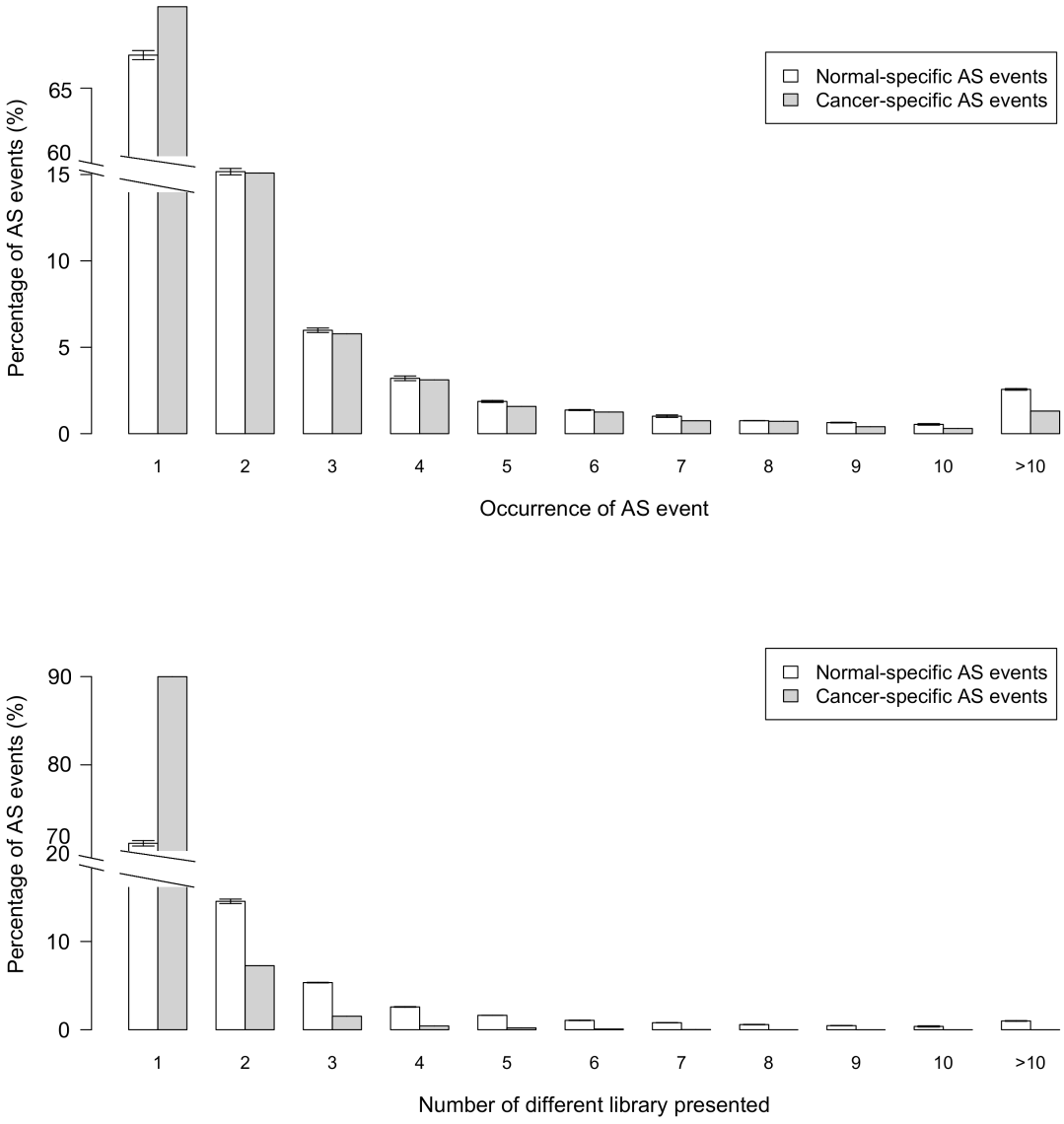


Figure 4. Normal and cancer specific AS events frequency distributions. Top panel shows the number of times each AS event is found and bottom panel shows the number of libraries where an AS event is found. Error bars in distributions from normal specific transcripts represent one hundred randomly selected samples from normal-specific transcripts of equal transcript and library number to the number of cancer-specific transcripts and libraries available.

5.3.2 Cancer transcripts show signatures consistent with splicing noise

We then assessed whether cancer libraries and in particular cancer-specific transcripts show signatures consistent with increased rates of splicing noise. If so, we

expect cancer transcripts to: A) have a higher incidence of nonsense or frameshift mutations which introduce a premature translation termination codons to mRNAs resulting in truncated proteins or more often rendering them vulnerable to nonsense mediated decay (NMD) (Green et al. 2003; Lewis et al. 2003). In the case of cancer-specific transcripts we can expect them to: B) have reduced identifiable functional components consistent with higher rates of aberrant incorporation of non coding regions into the transcript (see methods); C) be found mostly as single copy and D) be present in only one library thus not being part of the core cancer transcription profile as these are more likely to result from splicing errors (Zhang et al. 2009).

Transcripts were classified according to whether they contained AS events found in both normal and cancer tissues or unique to either resulting in four groups: 1) *Normal common*, with transcripts containing AS events also found in at least one cancer library, 2) *Normal-specific*, whose AS events are only found in normal tissue samples, 3) *Cancer common*, containing transcripts from cancer libraries with AS events also found in at least one normal tissue library and 4) Cancer-specific with transcripts with AS events unique to cancer libraries. Our results show, compared to normal tissue derived transcripts, an increased incidence of premature stop codons among cancer-derived transcripts which is higher for cancer-specific transcripts (Figure 2, $P < 0.0001$) in both human and mouse. In both species, cancer-specific events were also found to have a significantly lower number of identifiable functional components ($P < 0.0001$; Figure 3). In addition, we found that the vast majority (79.0%) have been sequenced only once with 90.5% identified in a single EST library in human (Figure 4). In contrast, normal-specific transcripts show less pronounced differences in premature stop codons and functional components compared to transcripts with normal-common AS events (Figure 2 and Figure 3). We also found that transcripts containing AS events particular to normal tissues are significantly less likely to be found as a single copy or confined to a single library ($P <= 0.0001$; Figure 4).

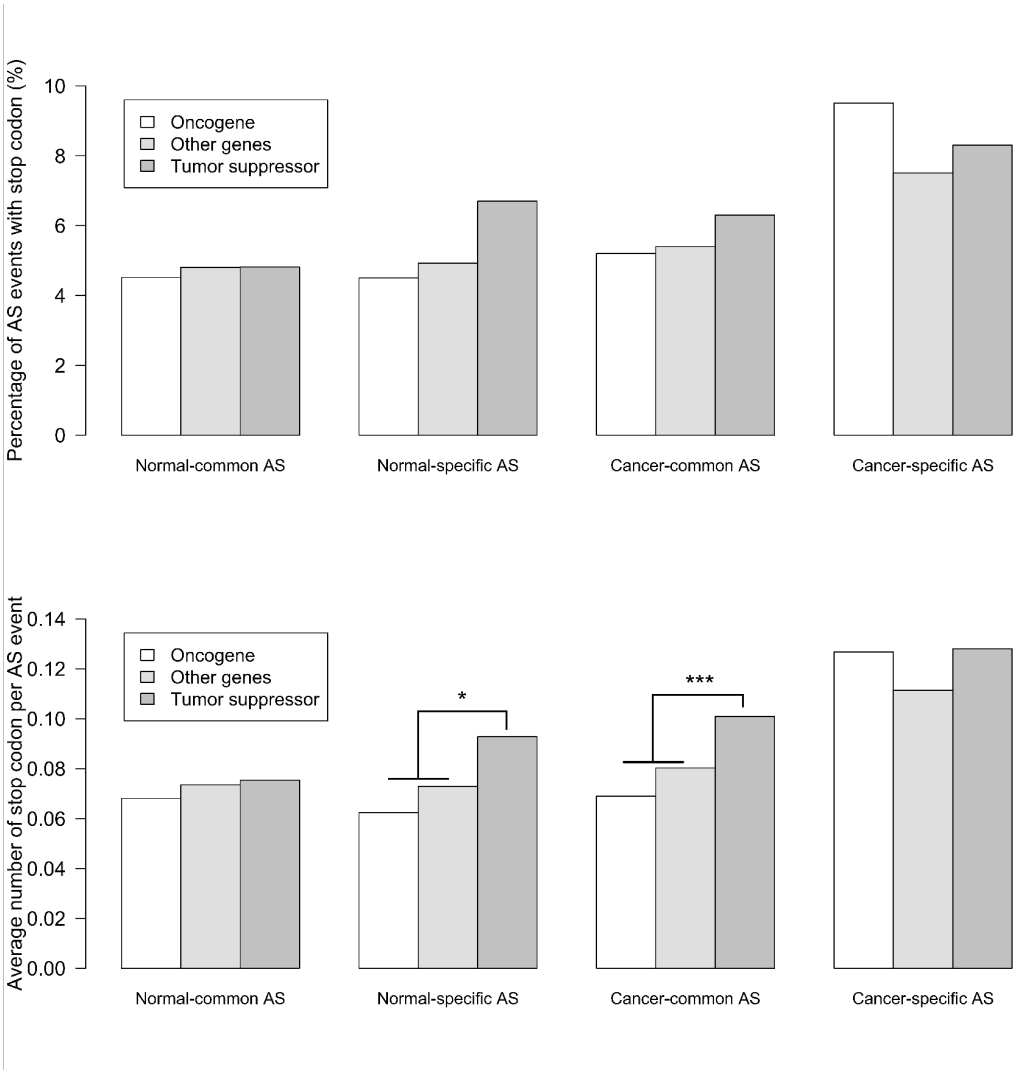


Figure 5. Premature stop codon frequency in oncogenes, tumour suppressor and *other* genes. Top panel shows the percentage of premature stop codon containing AS events. Bottom panel shows the average number of stop codons per AS events. AS events were classified depending on whether they were derived from oncogenes tumour suppressor and *other* genes. Broader groupings from Figure 2 and Figure 3 are also labelled. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with 0.01< P < 0.05 (*), 0.001< P< 0.01 (**) and P <= 0.001 (***).

5.3.3 Tumour suppressor and oncogenes reveal contrasting transcript quality reductions in cancer libraries

Because tumour suppressor and oncogenes play a key role in tumour progression, we tested whether these gene categories presented any differences in the frequencies of disabled transcripts. Inactivation of tumour suppressor genes *NFI*, *FHIT* and *TSG101* and strengthening oncogenes *CD44* and *RON* by AS have been reported (reviewed in

(Kalnina et al. 2005; Skotheim and Nees 2007). To test whether splicing noise signatures affect tumour suppressor and oncogenes differently, we divided all genes into oncogenes (648), tumour suppressor (850) and *other* genes according to the CancerGenes database (Higgins et al. 2007). We found that even if as a whole cancer-derived transcripts are more likely to contain premature stop codons consistent with misplicing (Figure 2), this increase is not equally distributed between gene categories (Figure 5). Common cancer-derived oncogene transcripts show only marginal changes in the rate of premature stop codons compared with transcripts derived from normal tissues (Figure 5). In contrast, tumour suppressor genes show a marked increase in the incidence of premature stop codons in cancer libraries (Figure 5, $P < 0.001$). These differences in transcript quality among gene categories are not observed in normal libraries.

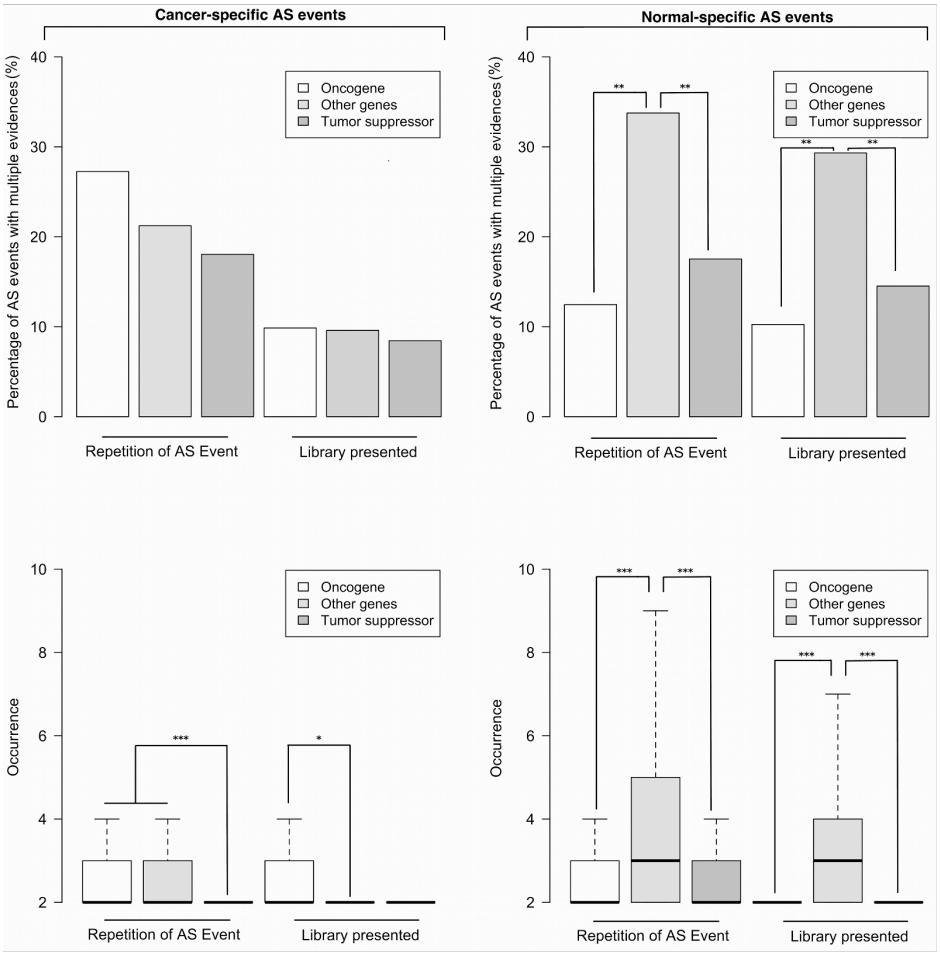


Figure 6. AS event frequency for normal and cancer transcripts divided into oncogene, tumour suppressor and *other* genes. Left and right panels represent cancer-specific and normal-specific AS events, respectively. Distributions for normal-specific AS events are the average results from 100 randomly selected samples of equal size to the number of cancer-specific AS events. Top panels present the percentage of AS events which are present in more than one copy and/or more

than one library. Bottom panels are a box plot of the average number of copies per AS event or the number of libraries where each AS event is present. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with 0.01< P < 0.05 (*), 0.001< P< 0.01 (**) and P <= 0.001 (***).

Analyses of transcripts specific to cancer or normal tissues showed that cancer-specific AS events have an elevated rate of premature stop codons in all three categories, further suggesting that a significant proportion of cancer-specific AS events containing transcripts are likely to result from splicing errors. We also found an elevated frequency in premature stop codons among tumour suppressor derived normal-specific AS transcripts (Figure 5; P = 0.016 and P = 0.014) which is not explained by the fact that these genes have a slightly longer average coding region (Figure 1S and 2S). When comparing transcript abundance in cancer-specific AS events (Figure 6), we found that oncogenes are more likely to produce cancer-specific AS events with more than one copy and to be found in more than one library than *other* genes (P = 0.037; Figure 6). This pattern is not found for normal-specific AS transcripts where the group of *other* genes were far more likely to be present in multiple copies and multiple libraries than both tumour suppressor and oncogenes (P < 0.0001; Figure 6).

In order to assess functional content, we examined the distribution of functional components for oncogenes, tumour suppressor and other genes in both cancer and normal AS transcripts. For alternative splicing events found in both cancer and normal libraries, oncogenes and tumour suppressor derived transcripts had higher frequencies of functional components compared to *other* genes (Figure 7, P = 0.008 and P = 0.04), suggesting that alternative splicing areas contribute significantly to the functional properties of these genes protein products. While among normal-specific AS areas there is a reduction in the functional content from oncogenes; in cancer-specific AS areas, it is tumour suppressor genes which show a marked reduction in functional content. No such reduction is observed among AS areas of oncogenes (Figure 7, P = 0.019).

5.4 Discussion

We have shown that transcripts derived from cancer libraries have an elevated rate of stop codons consistent with increased rates of missplicing in cancer transcriptomes. Transcripts with alternatively splicing events unique to cancer libraries showed an even

greater enrichment in premature stop codons (Figure 2) as well as containing fewer identifiable functional domains (Figure 3). Importantly, all cancer-specific transcripts were found in fewer than ten cancer libraries (out of a total of 367) with almost 80% of them found as a single copy (Figure 4). These features suggest that a significant proportion of these transcripts are unlikely to produce a functional protein product and given that no cancer specific transcripts was found to be ubiquitous to all cancer libraries or even a cancer type, we believe that the majority of cancer-specific transcripts, although probably functional, are unlikely to form part of a core cancer-transcriptome. Thus we estimate that the clinical and diagnostic relevance of particular cancer-specific transcripts may prove rather limited.

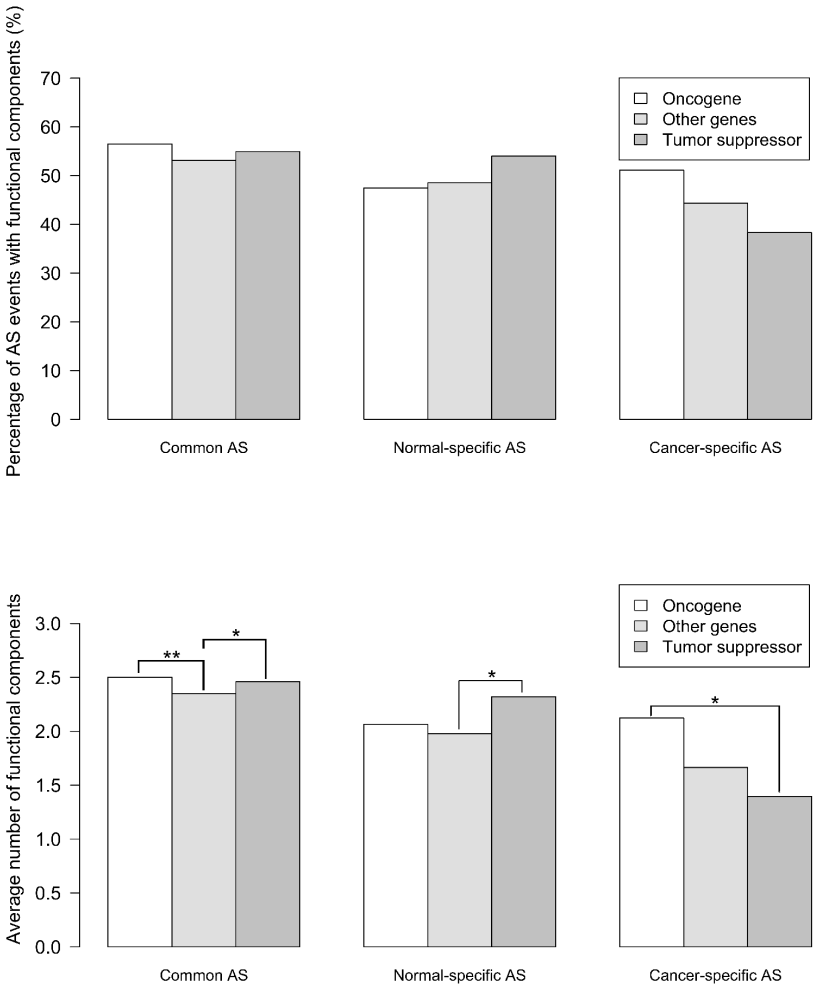


Figure 7. Identifiable functional components in AS events in cancer and normal transcripts divided into oncogene, tumour suppressor and *other* gene-derived. Top panel shows the percentage of AS events with at least one identifiable functional component (see methods). Bottom panel shows average number of identifiable functional components per AS area. In both panels, AS events were divided into groups as in Figure 3 and further subdivided into oncogene, tumour suppressor and

***other* genes. Stars represent significant differences among groups from top panels (Chi-square test) and bottom panels (Wilcoxon test) with 0.01< P < 0.05 (*), 0.001< P< 0.01 (**) and P <= 0.001 (***).**

In contrast, analyses of transcripts only found in normal tissue samples did not reveal a similar increase in noise signatures (Figure 2 and 3) and a significantly greater proportion were found in multiple libraries (Figure 4). Mutations leading to the absence of these transcripts in cancer libraries may have a role in cancer establishment and its progression and may therefore warrant further studies examining their clinical potential.

Interestingly, when dividing genes into oncogenes, tumour suppressors and *other* genes, we found marginal increases in stop codons in oncogene derived transcripts in cancer libraries while tumour suppressor genes showed a strong increase in premature stop codons. We found a higher incidence of premature stop codons amongst tumour suppressor genes in both normal-specific and cancer-common AS (Figure 5). This is not explained by differences in coding region length (Figure S1 and S2). The fact that cancer-specific oncogene transcripts have a higher functional content compared to those normal specific, suggests that, in some instances, oncogene-derived cancer-specific transcripts may confer novel functional properties to protein products potentially having a role in cancer cells. Given that this set of transcripts are mostly found in single libraries it is likely that their functional contribution is likely to be specific to cancers of individual patients.

We conclude that cancer states are associated with an elevated rate of aberrant transcripts particularly pronounced in tumour suppressor genes but from which oncogenes are spared. We therefore suggest that splicing noise should be considered when evaluating cancer-specific splicing events as they have a significant higher incidence of premature stop codons. Given that nonsense mutations affect only a minority of transcripts, it is feasible to assume that most cancer and normal specific transcripts may be transcribed into functional proteins and may contribute significantly to the cancerous phenotype. Nevertheless, the fact that most cancer-specific splice variants we identified are found as single copies in one EST library may somewhat limit their value as wide spectrum diagnostic probes and/or treatment targets. Assessment of global AS signatures by gene category may be more promising. Finally we propose that the roles of normal-specific and mutation in common alternative splicing variants should be examined in addition to cancer-specific transcripts; analyses of these absent AS transcripts may further aid in the understanding of the cancer physiology.

5.5 Supplementary Materials

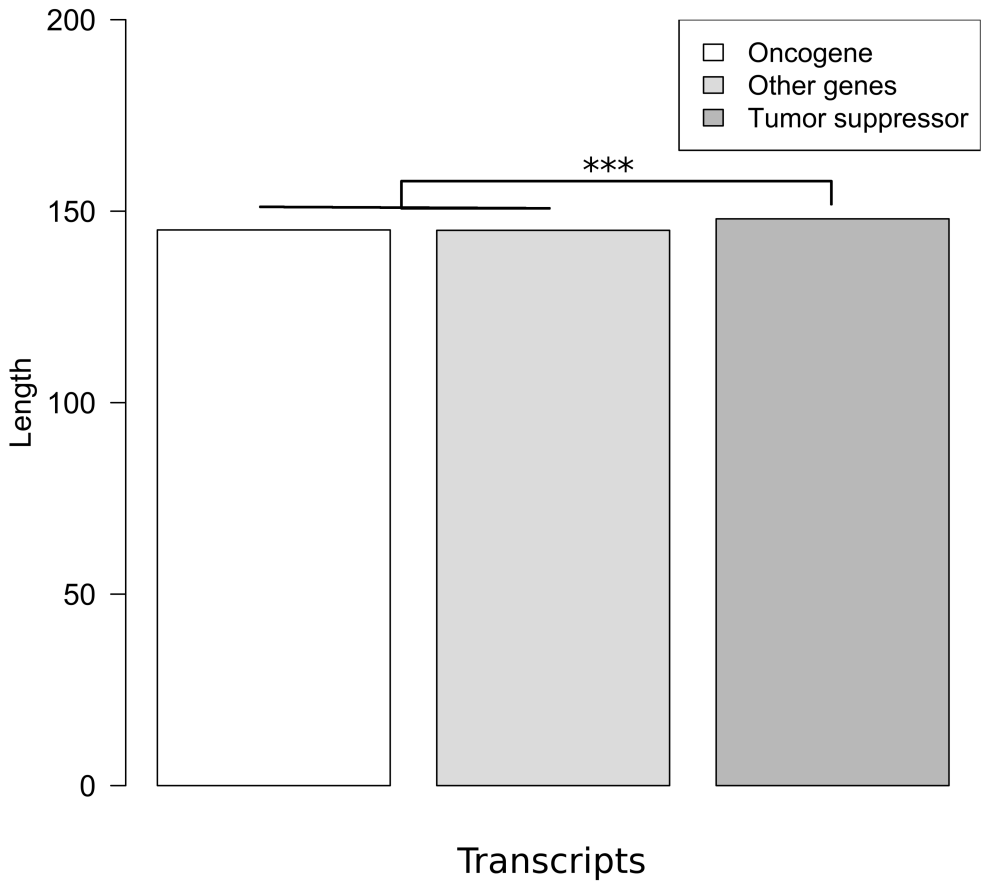


Figure S1: Comparison of the length of coding region among oncogene, other genes and tumour suppressor genes. The length of transcripts of tumour suppressor genes is significant longer than other genes and oncogene. Stars represent significant differences among groups (Wilcoxon test) with 0.01< P < 0.05 (*), 0.001< P< 0.01 (**) and P <= 0.001 (***).

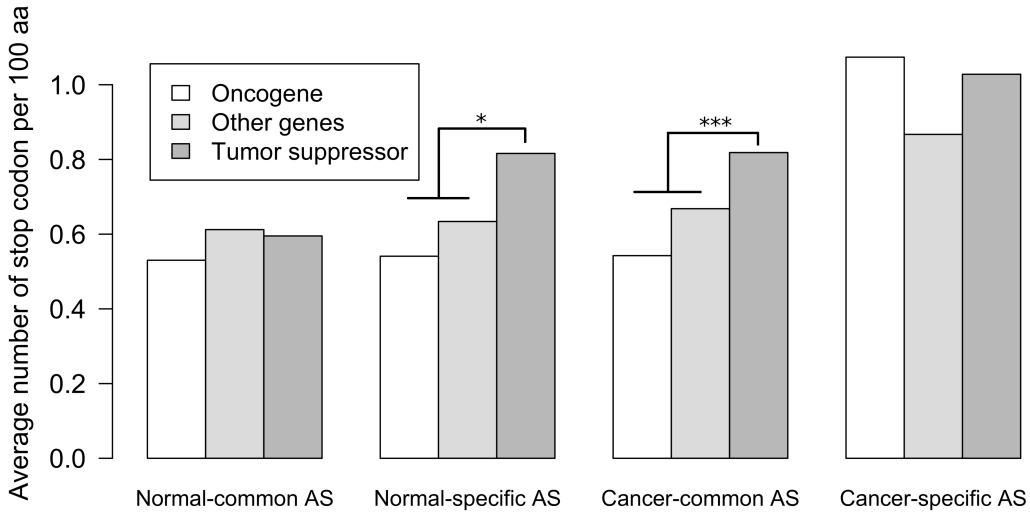


Figure S2: Average number of stop codons per 100 amino acids in oncogene, tumour suppressor and *other* genes. Broader groupings from Figure 2 and Figure 3 are also labelled. Stars represent significant differences among groups (Wilcoxon test) with 0.01< P < 0.05 (*), 0.001< P< 0.01 (**) and P <= 0.001 (***).

6 General discussion

Over a decade has passed since the publication of the first draft of the human genome. Since then, thanks to the development of increasingly powerful sequencing technologies, genomes of species from most major taxa have been sequenced. However, we still have a very limited understanding of how genes are controlled and how they underline observed phenotypes. For the last three years, I have focused my efforts on the study of alternative splicing as a potential source of functional innovation in the genome through evolution and disease states.

6.1 Alternative splicing database: ECCASED

Alternative splicing events detection is currently mostly based on transcript sequencing or exon based microarray technology. Genome sequence based prediction of alternative splicing events and prediction of levels of splicing remain unreliable. Although numerous alternative splicing datasets for multiple species are available (Kim et al. 2007b; Lee et al. 2007; Bhasi et al. 2009; Koscielny et al. 2009a), none takes into account the well documented fact that AS detection is highly dependent on transcript coverage for genes within a genome or between species (Brett et al. 2002; Kan et al. 2002; Kim et al. 2007a; Nilsen and Graveley 2010). The lack of comparable estimates of AS levels impose serious limitations on the interpretation of any analyses of AS levels and its relation to other genomic and phenotypic variables especially with expression levels or any parameter co-varying with it. In order to explore how alternative splicing has evolved through time I decided to construct a comparative resource for alternative splicing data. In Chapter 2, I presented the Eukaryotic Comprehensive & Comparable Alternative Splicing Events Database (ECCASED) based on the analyses of over 30 million ESTs for 114 eukaryotic genomes, including protists (22), plants (20), fungi (23), metazoan (non-vertebrates, 29) and vertebrates (20). Using a uniform analyses pipeline for all species the ECCASED database provides both comprehensive identification of AS events based on all available ESTs per gene and a comparable AS index using a random sampling protocol thereby overcoming biases in AS detection caused by differential transcript coverage. This data was then linked to a functional annotation including GO

terms and expression patterns of genes. Orthology annotations were used to allow comparative analysis among different species. The ECCASED database represents the first assessment of AS patterns for over 70% of the 114 species included in it. Most importantly, by providing an AS index not biased by transcript coverage differences, ECCASED is the first resource to provide a comparable AS index allowing direct comparisons for multiple genes within and between species.

Given its potential benefit for the wider scientific community a web interface has been built (http://bio.bdfield.com/eccased/index.php). At the moment the ECCASED database allows one to inspect all AS events detected for each gene in 114 species of eukaryotes. By providing a comparable measure of AS based in random samples the ECCASED database also allows to make direct comparisons of the levels of AS across genes within and between genomes. Over the next year I plan to further improve this database in three ways: First, I will use next generation sequences to complement longer reads from EST sequences. The shorter but more abundant RNA-seq reads will allow us to: 1) assess whether AS events detected in a single EST is likely to be a splicing error; 2) identify new AS events not found in EST sequences and 3) more accurately calculate the levels of expression of specific isoforms which contain unique AS events. Secondly, alignments of orthologous genes with AS annotations will allow to directly assess conservation of AS events. Thirdly, where data allows for it, I plan to add transcript abundance per tissue for each AS isoform.

6.2 How does alternative splicing correlate to gene duplication

Alternative splicing is a potential source of transcript diversity, in addition to gene duplication, it is relevant to explore the relationship between alternative splicing and gene family size. Recent studies had reported that in some eukaryotes alternative splicing is inversely correlated with gene family size suggesting that the two processes are to some extent mutually exclusive but it hints that an optimal transcript diversity per gene family exists and is dynamically maintained by either increased alternatively spliced isoform through time (Roux and Robinson-Rechavi 2011) or decreased AS after events of gene duplication (Su et al. 2006) or a combination of both. I believe that existing studies present a number of methodological flaws, most importantly, they fail to account for the differential EST coverage associated with high and low expression of genes resulting in a

strong association between EST coverage and detected splicing events. In Chapter 3, using a random sampling of a set number of ESTs per gene to identify AS events, we generated alternative splicing estimates which are comparable across different genes and species in 17 species from plants to mammals.

My results of the analysis of gene family expansion and AS levels in 17 species along the eukaryotic tree shows that there is no consistent trend of an inverse relationship between AS and GFS, and that in those species where the pattern is observed, the amount of variance explained is very small. Instead, the marginal relationship between AS and GFS in some species appears to be the result of the strong relationship between AS and GFS with breadth of expression. However, when looking at the patterns of transcript diversity increases on a per gene family basis across species, it is clear that for gene families with an expanding total number of distinct transcripts, AS and GFS are positively correlated. AS and GFS are negatively correlated only in gene families where there is no increase in overall transcript diversity. This suggests that AS and GFS do appear to be coupled processes which work together in diversifying gene families and become antagonistic in those gene families with a stable number of transcripts being produced suggesting that in those there is an optimum transcript diversity level.

6.3 Alternative splicing and gene duplication contributes to transcript diversity expansion in eukaryotes

One of the most fascinating consequences of the proliferation of genome sequences and transcript data in increasing number of species is the fact that it allows, for the first time, to examine the relationship between genomic features and observed phenotypes. One of the most elusive phenotypes is the evolution of complexity as how to measure this phenotype is controversial in itself. What actually underpins the complexity remains enigmatic. Taking the estimated number of cell types as a proxy of complexity, several studies have reported a link between complexity and various genomic features. There are two main explanations for the evolution of organism complexity. First is that the increasing regulation with the stable gene number. For example, the increasing complexity of gene regulation (Warnefors and Eyre-Walker 2011), protein-protein interaction domains on organism and network complexity (Xia et al. 2008a) or epigenetics factors such as non-coding RNAs and DNA methylation (Costa 2008).

Second one is by expanding the number of distinct proteins produced in an organism. However, these two hypotheses could be overlapped. Alternative splicing, as a product of the regulation of gene expression, will generate more proteins, which will then increase the interaction between proteins and form more complicated network. Therefore, I believe that multiple factors and changes in regulation levels contributed to the increasing complexity through evolution.

In Chapter 4, through the systematic analyses of over 27 million publicly available full and partial transcripts from 18 eukaryotic species, I provide evidence for a strong increase in alternative splicing over the last 1400 million years. Importantly, our proteome size estimates, ranging from ~7000 in yeast to ~90000 in human, closely covary with organism complexity –assayed as cell type number. Compared to genome size, gene number, non-coding DNA parameters and functional domain content previously shown to covary with complexity, proteome size is by far the strongest genomic-derived predictor of organism complexity, explaining over 70% of the variance. These results could suggest that proteome expansion fuelled by alternative splicing and gene duplication constitutes one of the fundamental components in the evolution of organism complexity.

6.4 Alternative splicing in cancer

Recent genome-wide analyses have detected numerous cancer-specific alternative splicing (AS) events (Wang et al. 2003; Xu 2003; Hui et al. 2004; Kim et al. 2008b; He et al. 2009). However, analysis of AS patterns in normal tissue derived transcript libraries have suggested that a significant proportion of detected AS events in fact result from instances of miss-splicing. Whether transcripts containing cancer-specific AS events are likely to be translated into functional proteins or simply reflect noisy splicing, thereby determining their clinical relevance, is not known. In Chapter 5, I show that consistent with a noisy splicing model, cancer-specific AS events generally tend to be rare, containing more premature stop codons and with less identifiable functional domains in human and mouse. Interestingly, common cancer-derived AS transcripts from tumour suppressor and oncogenes show marked changes in premature stop codon frequency differences with tumour suppressor genes exhibiting increased levels of premature stop codons whereas oncogenes have the opposite pattern. We conclude that tumours tend to have faithful oncogene splicing, a higher incidence of premature stop codons among

tumour suppressor and cancer-specific splice variants and suggest the importance of normal-specific and mutation in common alternative splicing variants.

This study proves that the present of novel transcripts in a disease state is not necessarily a reflection of an important biological role. Instead, novel transcript isoforms may in most cases reflect the fact that even under normal conditions, splicing regulation is not perfect. We believe that instances of novel transcript/proteins becoming integral to the cancer machinery are rare. However, I most stress that whether alternative splicing does play a key role in diseased transcriptomes including cancer remains unknown. Changes in the relative and absolute levels of non-aberrant alternative splicing variants compared to normal tissues might in fact be decisive in cancer progression and maintenance. In addition, it may well be that increases or decreases in AS noise itself on a genome wide basis or for a smaller subset also impacts on cell behaviour. As the increasing number of the complete DNA sequence of cancer genomes provide us a comprehensive perspective of how cancers have developed (Stratton et al. 2009; Meyerson et al. 2010), it will be possible to better assess whether changes in splicing regulation plays a significant role. If so, then simply measuring changes in overall expression levels per gene may turn out to be missing key information.

6.5 General conclusion

Alternative splicing as a source of novel transcripts without the need of gene duplication events has been seen as a potential key player in the evolution of the eukaryotic genome since the sequencing of the human genome revealed it to have around 3 times as many genes as the humble baking yeast. Despite the increasing interest in alternative splicing, however, little is known about AS levels in but a few species. The difficulties in predicting AS events from genomic sequences alone, together with the fact that AS event detection is strongly influenced by transcript coverage of a gene, has undoubtedly slowed down the study of how alternative splicing has evolved over time, how AS is regulated, and how it may relate to other genomic features and crucially to phenotype. To facilitate comparative genomics studies of AS, I implemented a consistent algorithm for transcript analyses in over 100 eukaryotic species. Comparable AS data allowed me to address a number of evolutionary questions regarding the evolution of AS

and its implications for the evolution of transcript diversity and complexity. I first assessed in Chapter 3 how AS relates to gene duplication events and found that contrary to recent studies carried out in a small number of species, AS and gene duplication are not negatively correlated on a genome-wide basis. Instead an inverse relation is only observed among gene families with a stable number of transcripts. AS has been proposed to be the missing source of complexity given that the number of genes in the human genome was well below expectations. My results presented in Chapter 4 appear to support to the notion that AS play a very important role in increasing transcript diversity and complexity in eukaryotes. Finally, in Chapter 5, by comparing normal and cancer tissue-derived transcript libraries I found that, contrary to previous suggestions, there is little evidence for cancer-unique AS transcripts playing an important role in cancer onset and progression as most cancer AS events were found to be single copy events and constrained to one library. Together my results provide some novel insights into the evolution of AS.

The analysis of AS is limited by the availability of transcript sequences. With the increasing popularity of next generation sequencing the study of alternative splicing is likely to undergo a revolution (Mortazavi et al. 2008). The higher depth of sequencing of transcriptomes in human and other species has increased our understanding of AS event expression patterns in different tissues (Wang et al. 2008; Kang et al. 2011), developmental stages (Graveley et al. 2011) and epigenetic regulation (Shukla et al. 2011).

This increasing amount of data is aiding the development of better AS event and tissue expression pattern predicting methods, further increasing our ability of performing comparative analysis of AS in species with no significant transcriptome coverage. Machine learning has been applied to predict the tissues-specific AS pattern in mouse (Barash et al. 2010). However, understanding the splicing code and regulation of AS, which are essential for predicting the AS pattern, will still be key issues given that regulation of AS occurs at many levels (Luco and Misteli 2011).

6.6 Future studies

I will now describe some ongoing projects in collaboration with both my PhD supervisor Dr. Urrutia and her group as well as with my postdoctoral supervisor Dr. Soranzo through which I expect to further my understanding of alternative splicing and its biological relevance.

1. Does alternative splicing impact gene duplicate retention?

Why do some duplicate genes survive while others are lost? After duplication, most extra gene copies soon accumulate disabling mutations and degrade, but some are retained. Over a dozen models have been proposed to explain fates of duplicate genes (reviewed in (Innan and Kondrashov 2010; Soskine and Tawfik 2010). One widely held model predicts that, after a duplication event, the accumulation of reciprocally complementary mutations disabling both duplicate copies of a gene plays an important role in the retention of duplicate genes with the immediate acquisition of a novel function being a rare event (Ohno 1970; Lynch and Conery 2000; Innan and Kondrashov 2010; Soskine and Tawfik 2010). There are no reported gene characteristics which would predispose any genes to be duplicated other than in cases where gene dosage increases are favorable. I hypothesize that alternative splicing may potentially increase the chances of subfunctionalisation events after duplication events since a relatively small number of substitutions may allow the reciprocal loss of AS isoforms in the two duplicate copies leading to subfunctionalisation. To test this, I decided to analyse gene duplicate retention after whole genome duplication events at the base of the vertebrate lineage. Preliminary analyses show that higher levels of ancestral alternative splicing in *Ciona* and amphioxus, significantly increases duplicate gene retention. Moreover, alternative spliced areas appear to have been reciprocally lost among duplicate copies. These observations suggest that alternative splicing shapes the survival chances of duplicate genes possibly by facilitating functional split between gene copies. If so, it would suggest that, at least in some cases, functional innovation and protein specialization precedes rather than arise from duplication events. In these cases, gene duplication would stabilise alternative splicing isoforms while facilitating the accumulation of novel splicing events in a cycle of increasing transcript innovation.

2. Are the increased levels of AS throughout evolution functionally relevant?

A number of studies have indicated that alternative splicing levels per gene have increased over time. My own analysis using a larger dataset of 18 species which corrects for differences in transcript coverage has also shown that AS has been an important contributor to transcript diversity expansion in the eukaryotic genome. Various reports however, suggest that a significant number of AS transcripts contain premature stop codons and have thus been proposed to be the result of splicing errors (Green et al. 2003; Lewis et al. 2003; Zhang et al. 2009). A comparison of AS patterns among orthologous genes found, however, that about 9% percent of conserved AS events between human and mouse result in premature stop codon-containing transcripts possibly adding an extra layer of regulation to gene expression (Mudge et al. 2011). Thus, whether the expansion in AS events over time is functional remains to be known. If alternative splicing increases along the phylogenetic tree are indeed functional we would expect these AS events to: A) have a lower incidence of internal stop codons (rendering them vulnerable to nonsense mediated decay (Green et al. 2003; Lewis et al. 2003)), B) have a relatively higher number of identifiable functional components consistent with lower rates of aberrant incorporation of non coding regions into the transcript (see methods); C) be found mostly as multi-copy rather than single copy occurrences; D) be present in more than one library and E) be conserved through evolution.

Preliminary data suggests that AS events have increased over time even when removing all transcripts containing premature stop codons or found in less than three copies. Moreover, the number of identifiable functional components per 100bp of transcript, either in AS regions or in the gene as a whole, has not decreased over time. Together these results suggest that even if a significant proportion of AS transcripts may indeed not be functional there is no indication that this proportion is growing, thus cancelling out the effects of transcript diversification.

3. Mapping genetic variation with alternative splicing in human genome and disease.

While expression QTL (eQTL) studies have illuminated the location and impact of genetic variants affecting gene expression, such analyses do not take into account the fact that unequal expression level exists among exons within a gene due to alternative splicing (AS). Recently, it has been reported that 94% human multi-exon genes undergo AS, and that up to 50% of the mutations that cause human disease may alter the efficiency and pattern of splicing. Therefore, AS may provide a potential mechanism underlying both phenotypic diversity and disease susceptibility in human populations. However, to what

extent genetic variation affects AS, how to infer splicing modification (e.g. splicing QTL) from genome-wide association studies (GWAS), and how to map genetic variation with alternative splicing pattern and identify candidate genes with disease-related splicing, remain largely unexplored. We used data from ~ 8 millions ESTs and ~40 millions SNPs from the 1000 Genome Project in humans, to assess the relationship between SNPs and AS. We found a strong positive correlation between the number of SNPs and AS occurrence, with SNPs being enriched in the splicing site recognition sequence, or AS regulation domains, of AS exons compared to exons that are constitutively transcribed ($p < 0.0001$). We further systematically searched 5,786 disease/trait-related SNPs previously identified through GWAS (http://www.genome.gov/gwastudies/), and identified 16 SNPs located in splicing sites and 129 SNPs in splicing regulatory motifs. Our results provide initial evidence of a possible important role of sQTL in modulating genetic traits and diseases. We suggest that further work improving the bioinformatic interpretation of sQTLs will be crucial to understanding of the splicing code. To achieve this goal, we propose the first comprehensive splicing map for each exon in the human genome, and a database of both reported and predicted sQTLs, which will facilitate future sQTL studies in human disease.

4. Alternative splicing in fungal species

While alternative splicing has been intensively characterised in human where up to 94% of multi-exon genes have been found to be alternatively spliced, little is known about this process in fungi. In baker's yeast (*S. cerevisiae*), the most highly studied fungal species, alternative splicing has only been reported in a few genes such as Src1 (Grund et al. 2008) and PTC7 (Juneau et al. 2009). Prevalence and patterns of alternative splicing across different fungal taxa remains unknown. Here we assess alternative splicing in 23 fungal species with sequenced genomes and over 30,000 EST transcript data. We are planning to investigate: 1) What is the prevalence of alternative splicing in fungal genes and variation across species? 2) What is the frequency of different types of alternative splicing to contrast with frequency? 3) How has alternative splicing evolved in different fungal lineages? 4) Are alternatively spliced transcripts in fungi functional?

7 References

Alt FW, Bothwell ALM, Knapp M, Siden E, Mather E, Koshland M, Baltimore D. 1980. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3′ ends. *Cell* **20**(2): 293-301.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17): 3389-3402.

Artamonova, II, Gelfand MS. 2007. Comparative genomics and evolution of alternative splicing: The pessimists' science. *Chemical Reviews* **107**: 3407-3430.

Ast G. 2004. How did alternative splicing evolve? *Nat Rev Genet* **5**(10): 773-782.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**(7294): 53-59.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL et al. 2004. The Pfam protein families database. *Nucleic Acids Research* **32**: D138-D141.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* **340**(4): 783-795.

Berget SM, C. M, Sharp PA. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA* **74**(8): 3171-3175.

Berglund AC, Sjolund E, Ostlund G, Sonnhammer ELL. 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Research* **36**: D263-D266.

Bhasi A, Philip P, Sreedharan VT, Senapathy P. 2009. AspAlt: A tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes. *Genomics* **94**(1): 48-54.

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res* **20**(2): 180-189.

Blencowe BJ. 2006. Alternative splicing: new insights from global analyses. *Cell* **126**(1): 37-47.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST--database for "expressed sequence tags". *Nat Genet* **4**(4): 332-333.

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. *Nat Genet* **30**(1): 29-30.

Brinkman BM. 2004. Splice variants as cancer biomarkers. *Clinical biochemistry* **37**(7): 584-594.

Chacko E, Ranganathan S. 2009. Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. *BMC Genomics* **10**.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**(11): 741-754.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**(1): 1-8.

Costa FF. 2008. Non-coding RNAs, epigenetics and complexity. *Gene* **410**(1): 9-17.

Crollius HR, Jaillon O, Bernot A, Dasilva C, Bouneau L, Fischer C, Fizames C, Wincker P, Brottier P, Quetier F et al. 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat Genet* **25**(2): 235-238.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**(10): 1700-1708.

Early P, Rogers J, Davis M, Calame K, Bond M, Wall R, Hood L. 1980. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**(2): 313-319.

Farre D, Alba MM. 2010. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Molecular Biology and Evolution* **27**(2): 325-335.

Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM. 2006. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome Biol* **7**(10): R89.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics* **17**(2): 100-107.

-. 2009. Alternative splicing regulation without regulators. *Nature structural & molecular biology* **16**: 13-15.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW et al. 2011. The developmental transcriptome of Drosophila melanogaster. *Nature* **471**(7339): 473-479.

Green RE, Lewis BP, Hillman RT, Blanchette M, Lareau LF, Garnett AT, Rio DC, Brenner SE. 2003. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics* **19**: i118-i121.

Grund SE, Fischer T, Cabal GG, Antunez O, Perez-Ortin JE, Hurt E. 2008. The inner nuclear membrane protein Src1 associates with subtelomeric genes and alters their regulated gene expression. *Journal of Cell Biology* **182**(5): 897-910.

Haider S, Ballester B, Smedley D, Zhang JJ, Rice P, Kasprzyk A. 2009. BioMart Central Portal-unified access to biological data. *Nucleic Acids Research* **37**: W23-W27.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C et al. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**: D258-D261.

Hartmann B, Castelo R, Minana B, Peden E, Blanchette M, Rio DC, Singh R, Valcarcel J. 2011. Distinct regulatory programs establish widespread sex-specific alternative splicing in Drosophila melanogaster. *Rna* **17**(3): 453-468.

Hawkins RD, Hon GC, Ren B. 2010. Next-generation genomics: an integrative approach. *Nat Rev Genet* **11**(7): 476-486.

He C, Zhou F, Zuo Z, Cheng H, Zhou R. 2009. A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis. *PloS one* **4**(3): e4732-e4732.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**(23): 2971-2972.

Heebal Kim RK, Jacek Majewski, Jurg Ott,. 2004. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet* **36**: 915-917.

Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, Welsh-Bohmer KA, Hulette CM, Denny TN, Goldstein DB. 2008. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol* **6**(12): e1.

Higgins ME, Claremont M, Major JE, Sander C, Lash AE. 2007. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic acids research* **35**(Database issue): D721-726.

Hughes AL, Friedman R. 2008. Alternative splicing, gene duplication and connectivity in the genetic interaction network of the nematode worm Caenorhabditis elegans. *Genetica* **134**(2): 181-186.

Hughes TA. 2006. Regulation of gene expression by alternative untranslated regions. *Trends in genetics : TIG* **22**(3): 119-122.

Hui L, Zhang X, Wu X, Lin Z, Wang Q, Li Y, Hu G. 2004. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* **23**(17): 3013-3023.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* **14**(10A): 1870-1879.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* **11**(2): 97-108.

Irimia M, Rukov JL, Penny D, Garcia-Fernandez J, Vinther J, Roy SW. 2008. Widespread evolutionary conservation of alternatively spliced exons in Caenorhabditis. *Molecular Biology and Evolution* **25**(2): 375-382.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**(7011): 946-957.

Jensen CJ, Oldfield BJ, Rubio JP. 2009. Splicing, cis genetic variation and disease. *Biochem Soc Trans* **37**(Pt 6): 1311-1315.

Jin L, Kryukov K, Clemente JC, Komiyama T, Suzuki Y, Imanishi T, Ikeo K, Gojobori T. 2008. The evolutionary relationship between gene duplication and alternative splicing. *Gene* **427**(1-2): 19-31.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**(5653): 2141-2144.

Juneau K, Nislow C, Davis RW. 2009. Alternative splicing of PTC7 in *Saccharomyces cerevisiae* determines protein localization. *Genetics* **183**(1): 185-194.

Kalnina Z, Zayakin P, Silina K, Line A. 2005. Alterations of pre-mRNA splicing in cancer. *Genes Chromosomes & Cancer* **42**(4): 342-357.

Kan ZY, States D, Gish W. 2002. Selecting for Functional Alternative Splices in ESTs. *Genome Res* **12**: 1837-1845.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G et al. 2011. Spatio-temporal transcriptome of the human brain. *Nature* **478**(7370): 483-489.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y et al. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**(7145): 714-719.

Kim E, Goren A, Ast G. 2008a. Alternative splicing: current perspectives. *Bioessays* **30**(1): 38-47.

Kim E, Goren A, Ast G. 2008b. Insights into the connection between cancer and alternative splicing. *Trends in genetics : TIG* **24**(1): 7-10.

Kim E, Magen A, Ast G. 2007a. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research* **35**(1): 125-131.

Kim N, Alekseyenko AV, Roy M, Lee C. 2007b. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Research* **35**: D93-D98.

Kopelman NM, Lancet D, Yanai I. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37**(6): 588-589.

Koscielny G, Le Texier V, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, Stanley E, Fallsehr C, Hofmann O, Kull M et al. 2009a. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics* **93**(3): 213-220.

Koscielny G, Le Texier V, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, Stanley E, Fallsehr C, Hofmann O, Kull M et al. 2009b. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics* **93**(3): 213-220.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* **305**(3): 567-580.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lee Y, Kim B, Shin Y, Nam S, Kim P, Kim N, Chung WH, Kim J, Lee S. 2007. ECgene: an alternative splicing database update. *Nucleic Acids Research* **35**: D99-D103.

Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31**(2): 180-183.

Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**(1): 189-192.

Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. *Nat Rev Genet* **4**(11): 865-875.

Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R. 2005. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Letters* **579**(9): 1900-1903.

Lu Z-X, Jiang P, Xing Y. 2011. Genetic variation of pre-mRNA alternative splicing in human populations. *Wiley Interdisciplinary Reviews: RNA*: 120.

Luco RF, Misteli T. 2011. More than a splicing code: integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Curr Opin Genet Dev*.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science* **327**(5968): 996-1000.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494): 1151-1155.

-. 2003. The origins of genome complexity. *Science* **302**(5649): 1401-1404.

Malko DB, Makeev VJ, Mironov AA, Gelfand MS. 2006. Evolution of exon-intron structure and alternative splicing in fruit flies and malarial mosquito genomes. *Genome Res* **16**(4): 505-509.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. *Nat Rev Genet* **12**(10): 671-682.

Martinez O, Reyes-Valdes MH, Herrera-Estrella L. 2010. Cancer reduces transcriptome specialization. *PLoS One* **5**(5): e10398.

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**(10): 685-696.

Mollet IG, Ben-Dov C, Felicio-Silva D, Grosso AR, Eleuterio P, Alves R, Staller R, Silva TS, Carmo-Fonseca M. 2010. Unconstrained mining of transcript data reveals

increased alternative splicing complexity in the human transcriptome. *NUCLEIC ACIDS RESEARCH* **38**(14): 4740-4754.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.

Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J. 2011. The origins, evolution and functional potential of alternative splicing in vertebrates. *Molecular Biology and Evolution*.

Nagaraj SH, Gasser RB, Ranganathan S. 2007. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform* **8**(1): 6-21.

Nakai K, Horton P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences* **24**(1): 34-35.

Nelson CE, Hersh BM, Carroll SB. 2004. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol* **5**(4).

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**(7280): 457-463.

Ogasawara O, Otsuji M, Watanabe K, Iizuka T, Tamura T, Hishiki T, Kawamoto S, Okubo K. 2006. BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Research* **34**: D628-D631.

Ohno S. 1970. *Evolution by gene duplication*. Springer, New York, New York.

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**(2): 87-98.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**(12): 1413-1415.

Parker CJ, Shawcross SG, Li H, Wang QY, Herrington CS, Kumar S, MacKie RM, Prime W, Rennie IG, Sisley K et al. 2004. Expression of PAX 3 alternatively spliced transcripts and identification of two new isoforms in human tumors of neural crest origin. *Int J Cancer* **108**(2): 314-320.

Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet* **6**(12).

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ et al. 2010. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**(11): 909-912.

Roth G, Dicke U. 2005. Evolution of the brain and intelligence. *Trends in Cognitive Sciences* **9**(5): 250-257.

Roux J, Robinson-Rechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res* **21**(3): 357-363.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S et al. 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **38**: D5-D16.

Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S et al. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **37**: D5-D15.

Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**(7371): 74-79.

Skotheim RI, Nees M. 2007. Alternative splicing in cancer: noise, functional, or systematic? *The international journal of biochemistry & cell biology* **39**(7-8): 1432-1449.

Soskine M, Tawfik DS. 2010. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* **11**(8): 572-582.

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* **344**: 1-20.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* **458**(7239): 719-724.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**(16): 6062-6067.

Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Res* **16**(2): 182-189.

Takeda Ji, Suzuki Y, Sakate R, Sato Y, Seki M, Irie T, Takeuchi N, Ueda T, Nakao M, Sugano S et al. 2008. Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *NUCLEIC ACIDS RESEARCH* **36**(20): 6386-6395.

Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res* **13**(10): 2260-2264.

Valentine JW, Collins AG, Meyer CP. 1994. Morphological complexity increase in metazoans. *Paleobiology* **20**(2): 131-142.

Venables JP. 2006. Unbalanced alternative splicing and its significance in cancer. *BioEssays : news and reviews in molecular, cellular and developmental biology* **28**(4): 378-386.

Venables JP, Klinck R, Koh C, Gervais-Bird J, Bramard A, Inkel L, Durand M, Couture S, Froehlich U, Lapointe E et al. 2009. Cancer-associated regulation of alternative splicing. *Nature structural & molecular biology* **16**(6): 670-676.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Vogel C, Chothia C. 2006. Protein family expansions and biological complexity. *Plos Computational Biology* **2**(5): 370-382.

Wang BB, Brendel V. 2006. Genomewide comparative analysis of alternative splicing in plants. *Proc Natl Acad Sci U S A* **103**(18): 7175-7180.

Wang ET, Sandberg R, Luo SJ, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**(7221): 470-476.

Wang G-S, Cooper TA. 2007a. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**(10): 749-761.

Wang GS, Cooper TA. 2007b. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**(10): 749-761.

Wang Z, Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna* **14**(5): 802-813.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**(1): 57-63.

Wang Z, Lo HS, Yang H, Gere S, Hu Y, Buetow KH, Lee MP. 2003. Computational Analysis and Experimental Validation of Tumor-associated Alternative RNA Splicing in Human Cancer. *Cancer Research*: 655-657.

Warnefors M, Eyre-Walker A. 2011. The accumulation of gene regulation through time. *Genome Biol Evol* **3**: 667-673.

Watson PM, Watson DK. 2010. Alternative Splicing in Prostate and Breast Cancer. *Cancer*: 62-76.

Wegmann D, Dupanloup I, Excoffier L. 2008. Width of Gene Expression Profile Drives Alternative Splicing. *Plos One* **3**(10).

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**(9): 1859-1875.

Xia K, Fu Z, Hou L, Han JD. 2008a. Impacts of protein-protein interaction domains on organism and network complexity. *Genome Res* **18**(9): 1500-1508.

Xia K, Fu Z, Hou L, Han JDJ. 2008b. Impacts of protein-protein interaction domains on organism and network complexity. *Genome Res* **18**(9): 1500-1508.

Xing Y, Lee C. 2006. Alternative splicing and RNA selection pressure — evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* **7**(7): 499-509.

Xu Q. 2003. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research* **31**(19): 5635-5643.

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M et al. 2011. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**(7367): 64-69.

Yu Y, Maroney PA, Denker JA, Zhang XH, Dybkov O, Luhrmann R, Jankowsky E, Chasin LA, Nilsen TW. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. *Cell* **135**(7): 1224-1236.

Zdobnov EM, Apweiler R. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**(9): 847-848.

Zhang Z, Xin D, Wang P, Zhou L, Hu L, Kong X, Hurst LD. 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol* **7**: 23.